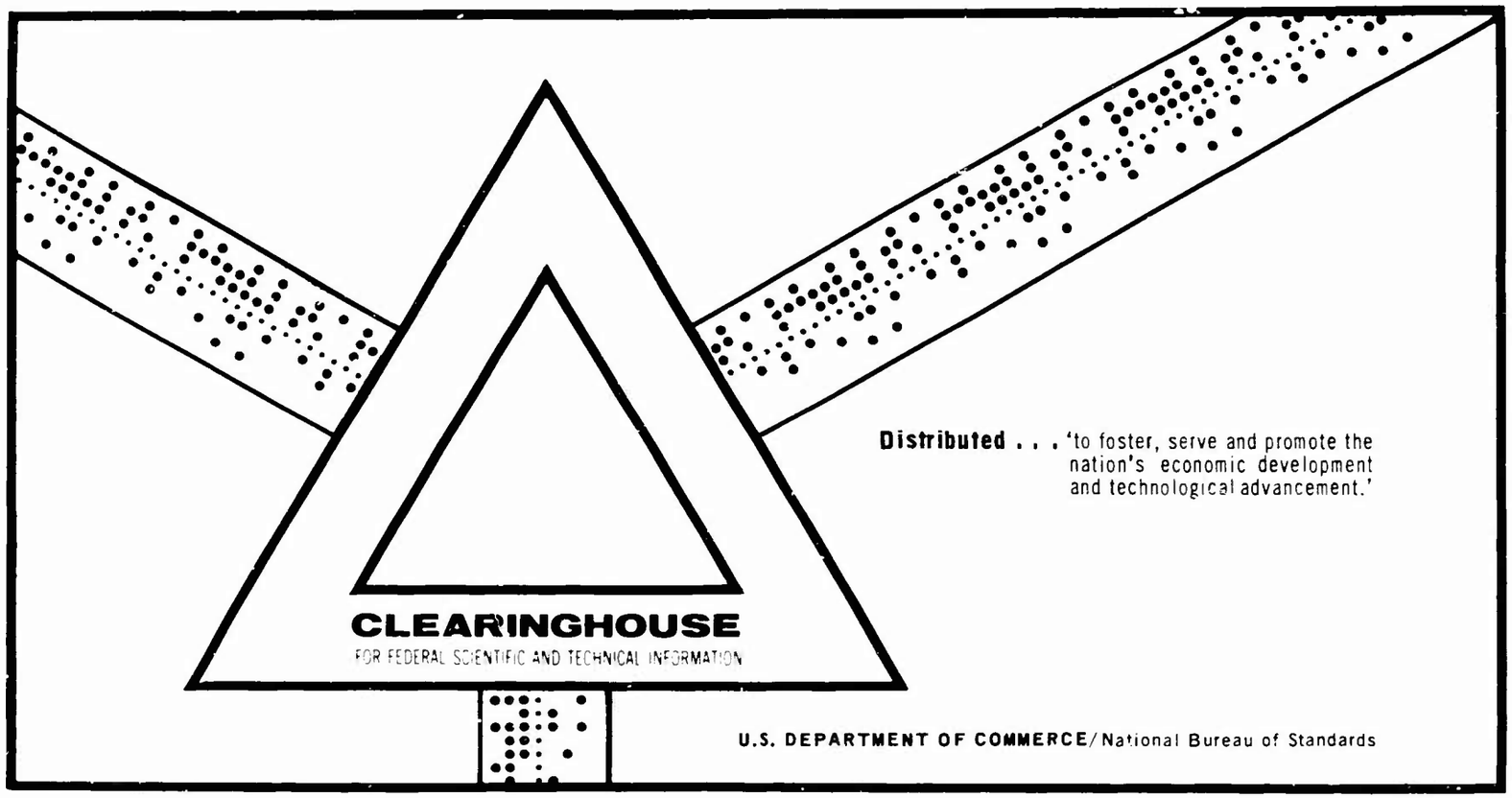
A MULTIPLE-DECISION APPROACH TO THE SELECTION OF THE BEST SET OF PREDICTOR VARIATES

John Schmidt Ramberg

Cornell University
Ithaca, New York

July 1969



DEPARTMENT OF OPERATIONS RESEARCH

COLLEGE OF ENGINEERING
CORNELL UNIVERSITY
ITHACA, NEW YORK 14850

DEPARTMENT OF OPERATIONS RESEARCH
COLLEGE OF ENGINEERING
CORNELL UNIVERSITY
ITHACA, NEW YORK

TECHNICAL REPORT NO. 79

July 1969

A MULTIPLE-DECISION APPROACH TO THE SELECTION OF THE BEST SET OF PREDICTOR VARIATES

by

John Schmidt Ramberg

Prepared under contracts
DA-31-124-ARO-D-474, U.S. Army Research Office-Durham
and
Nonr-401(53), Office of Naval Research

Table of Contents

1. Introduction and summary

In this thesis we will deal with selection procedure problems involving multivariate normal populations. These problems are of two distinct types: The first involves k bivariate normal populations, where the goal is to select the "best" population. The second involves one (k+1)-variate normal population, where the goal is to select the "best" set of a preassigned number of variates. Our major interest will center on problems of prediction and hence the "best" population or "best" set of variates will be determined on the basis of measures of "goodness" such as the simple population correlation coefficient and the multiple population correlation coefficient. We will limit our consideration to single-stage procedures.

The problem of selecting a variate or set of variates for prediction of a designated variate occurs in many areas. Hotelling [34] formulated such a problem as a hypothesis testing problem. We formulate this problem and similar but more general ones as multiple-decision problems. Our major objective is to provide a rational basis for determining the sample size necessary to insure that the probability of correctly selecting the "best" population (or set of variates) is sufficiently high whenever the "best" population (or set of variates) is better than the "next best" population (or set of

variates) by at least a predetermined amount.

1.1 The literature

A considerable literature on selection procedures already exists. This discussion is not intended to be a complete review of the literature of multiple-decision selection procedures, but rather to provide the reader with some idea of the general nature of the previous work which is relevant to this thesis.

Although there are many ways in which selection procedures can be formulated, the two most common formulations in the literature are the "indifference zone" approach as proposed by Bechhofer [5] and the "subset" approach as proposed by Gupta [27].

Bechhofer [5] formulated and solved the "indifference zone" approach for the problem of ranking the means of k univariate normal populations with common known variances, employing a single-stage procedure. Bechhofer, Kiefer, and Sobel [9] have written a monograph, Sequential Identification and Ranking Procedures, in which they discuss sequential solutions to the problems of ranking parameters of Koopman-Darmois populations (the normal means problem thus being considered as a special case). This monograph also contains a very complete bibliography on multiple-decision selection and ranking

procedures (including both the "indifference zone" and the "subset" approach). Paulson [42] has proposed a sequential procedure for the normal means problem which is quite different from that described in [9].

Bechhofer, Dunnett, and Sobel [7] gave a two-stage procedure for the normal means problem when the common variance is unknown and more recently Robbins, Sobel and Starr [46] have given a sequential solution to this problem. Several other papers have considered the "indifference zone" formulation for other univariate distributions and/or other parameters. The paper on ranking variances of univariate normal populations by Bechhofer and Sobel [10] is of particular relevance to this thesis.

A similar development has occurred for the "subset" formulation of the ranking problem. We will be concerned only with the "indifference zone" approach in this thesis and refer the reader to Gupta [27] for a description of the "subset" approach.

While most of the previous work has dealt with univariate populations, some recent papers (Alam and Rizvi [1], Gnanadesikan [23], Gupta [28], Gupta and Panchapekesan [29],[†]

[†]During the period in which this thesis was being written in final form, this paper was delivered at the Second International Symposium on Multivariate Analysis. The problem that we consider in Section 2.3 is a special case of the one described in this paper.

Gupta and Studden [30], Krishnaiah [36], Krishnaiah and Rizvi [37], and Thornby [49]), using a variety of approaches, have considered single-stage procedures for ranking problems involving k p-variate ($k \geq 2$, $p \geq 2$) normal populations. Here unlike the problems involving univariate populations, the vectors or matrices of parameters do not have a single "natural" ranking, but can be ranked according to many different criteria. All of these papers consider problems wherein the "goodness" of a population is measured in terms of predetermined univariate functions of its parameters. The problems considered in these papers can be categorized on the basis of these functions. One category consists of those problems in which this function is the Mahalanobis distance $\underline{\mu}_i \Sigma_i^{-1} \underline{\mu}_i'$, where $\underline{\mu}_i$ and Σ_i are the population mean vector and the population covariance matrix, respectively, of the i^{th} p-variate population, and includes problems considered in [1], [28], [30], and [37]. Another category consists of those problems in which this function depends only on the elements of the $\underline{\mu}_i$ and includes problems considered in [23], [37], and [49]. A third category consists of those problems in which this function depends only on the elements of the Σ_i and includes problems considered by [23], [29], and [36]. The problems considered in Part I of this thesis fall into this third category.

The problems of all three of these categories appear to

be rather artificial and are mainly of theoretical interest. In this thesis our principal reason for considering problems of this nature is that they provide some insight into the solutions of problems which do have a practical significance. These latter problems which are associated with a single multivariate normal population are considered in Part II of this thesis.

Several papers have also been written on another class of problems--ranking the variates of a single p-variate population. There are numerous possible "natural" rankings of the variates for this class of problems, including the "natural" ranking for the corresponding problem involving k univariate populations. These problems are complicated by the correlation structure (which may be unknown) between these variates, since most of the general theorems concerning ranking (e.g , see Barr and Rizvi [3]) assume the variates to be independent.

Bechhofer, Elmaghraby and Morse [8] have given a single-stage solution for the problem of selecting the variate with the largest single-trial cell probability for a single multinomial population; a sequential solution has also been given by Bechhofer, Kiefer, and Sobel [9].

Gnanadesikan [23] has given a single-stage solution for the "subset" approach to the problem of selecting the variate with the largest standardized population mean for both the

case of a known and an unknown covariance matrix. Bechhofer, Kiefer and Sobel [9] have shown that their sequential solution for Koopman-Darmois populations holds for the problem of selecting the variate with the largest population mean from a single multivariate normal population with a known covariance matrix (σ_{ij}) of the form $\sigma_{ij} = \begin{cases} \sigma & i=j \\ \rho\sigma & i\neq j \end{cases}$ and also for the problem of selecting the variate with the largest population variance from a single bivariate normal population with known or unknown population means and known population correlation coefficient. The problems considered in Part II fall within this general class of problems, which involve ranking the variates of a single multivariate population.

1.2 Summary

This thesis consists of two distinct, but related parts. In both parts our ultimate objective will be to provide a rational basis for determining the sample size for an experiment in which the goal is to select the "best" population (or set of variates) when certain probability requirements (which will be described precisely in later sections) are to be guaranteed. The formulation that we adopt here falls within the framework of the "indifference zone" ranking approach as proposed by Bechhofer [5]. Many of the ideas considered in this thesis could also be carried over to the "subset"

formulation of Gupta [27], but we do not consider his approach here.

In order to determine the exact sample size required for the "indifference zone" approach, one must know the exact joint distribution of the statistics on which the decision procedure is based. For many of the problems which we consider in this thesis, in particular those in Part II, the determination of these exact distributions appears almost hopeless. Hence we will find the asymptotic joint distribution of these statistics and work with it in the same manner as one would with the exact joint distribution were it known. That is, we will obtain the infimum of the asymptotic probability of correct selection over the region of preference for a correct selection and choose our sample size in such a way that this infimum satisfies the probability requirement. For certain of these cases the exact infimum of the asymptotic probability, referred to above, is also difficult to obtain, and in these cases we find a lower bound to this asymptotic probability and obtain the infimum of this lower bound. Using these results, we find a conservative approximation to the asymptotic sample size necessary to satisfy the probability requirement.

In Part I we consider problems concerning the selection of a "best" population from a set of k independent bivariate normal populations. We will be interested in attempting to

predict one of the variates of the bivariate normal population on the basis of the other variate. The "goodness" of this prediction will be measured in terms of three different criteria--the population conditional variance, the algebraic value of the population correlation coefficient, and the absolute value of the population correlation coefficient.

All three of the problems considered in this part are rather artificial and are mainly of theoretical interest. Our principal reason for studying them is that they provide some insight into similar, but more difficult problems (which do have practical importance) considered in Part II.

In Part II we consider the problem of selecting the "best" set of a preassigned number t of variates from a set of k ($t < k$) variates (which we term the predictor variates) for predicting a designated variate (which we term the predictand) in a (k+1)-variate normal population. Throughout this part the "best" set of predictor variates will be defined to be that set of t predictor variates for which the predictand has the smallest population conditional variance or equivalently that set of t predictor variates with which the predictand has the largest population multiple correlation coefficient.

In Chapter 3 we give the formal problem statement in its most general form along with the notation and an expression for the asymptotic probability of correct selection for this

problem.

Throughout Chapter 4 we assume that the predictor variates are uncorrelated and seek the asymptotic sample size which satisfies the probability requirement. For ($k = 2$, $t = 1$) we accomplish this objective by finding the infimum of the asymptotic probability of correct selection. For $k > 2$ we obtain approximations to the asymptotic sample size by finding lower bounds on the asymptotic probabilities of correct selection and then obtaining lower limits for these bounds over the region of preference for a correct selection. For ($k > 2$, $t = 1$) we use the Slepian inequality (Appendix B) to obtain this lower bound on the asymptotic probability of correct selection and for ($k > 2$, $t > 1$) we use the Bonferroni inequality (Appendix B) to obtain this lower bound.

In Chapter 5 we drop the assumption of uncorrelated predictor variates and proceed in a manner similar to that described in Chapter 4, restricting consideration to the case $t = 1$. For ($k = 2$, $t = 1$) we obtain the exact asymptotic sample size which satisfies the probability requirement. For ($k > 2$, $t = 1$) we use the Bonferroni inequality to obtain a lower bound on the probability of correct selection and then find the infimum of this lower bound over the region of preference for a correct selection and the corresponding approximation to the asymptotic sample size.

In Chapter 6, we suggest some areas of future research.

The asymptotic distribution theory of the sample statistics used in the decision procedures of Part II is given in Appendix A. The Bonferroni and the Slepian inequalities are given in Appendix B. In Appendix C two sample size approximations derived from these inequalities for the problem of ranking the means of normal populations with common known variances are compared with the exact sample size.

PART I

k Bivariate Normal Populations

In Part I we will consider situations in which we have k independent random vectors[†] $\underline{x}_i = (x_{i1}, x_{i2})$, each of which has a bivariate normal distribution with unknown mean vector $\underline{\mu}_i = (\mu_{i1}, \mu_{i2})$ and unknown covariance matrix Σ_i, where the elements of Σ_i are given by $\sigma_{i;rs}$ $(r,s = 1,2)$. Throughout this part we will be interested in attempting to predict one of the variates of a bivariate normal population on the basis of the other variate. The "goodness" of the prediction will be measured in terms of the population conditional variance, the algebraic value of the population correlation coefficient, and the absolute value of the population correlation coefficient.

Our objective will be to provide a rational basis for determining the sample size for an experiment in which the goal is to select the "best" bivariate population when certain probability requirements (which will be described in later sections) are to be guaranteed. The formulations that we will adopt fall within the framework of the so-called "indifference zone" ranking approach as proposed by Bechhofer [5].

[†]For simplicity, we will not distinguish notationally between random variables and their observed values.

2. Three formulations

We consider the three formulations of Part I in this chapter. For the first formulation we use the exact joint distribution of the statistics on which the decision proc is based, in order to determine the minimum sample si:e will guarantee the probability requirement.

For the latter two formulations we find the joint asymptotic distribution of a transformation of the statistics used in the decision procedures, and work with this asymptotic distribution in the same manner as one would with the exact distribution. That is, we obtain the infimum of the asymptotic probability of correct selection over the region of preference for a correct selection and then choose our sample size in such a way that this infimum satisfies the probability requirement.

2.1 Conditional variance formulation

In Section 2.1 we will be concerned with the problem of selecting the "best" population from a set of k independent bivariate normal populations Π_i (i = 1, 2, ..., k), the "goodness" of the Π_i being measured in terms of the population conditional variances

$$(2.1) \qquad \mathrm{Var}(x_{i1}|x_{i2}) = \sigma^2_{i;1|2}.$$

We denote the ranked values of the $\sigma^2_{i;1|2}$ by

$$\sigma^2_{[1];1|2} \leq \sigma^2_{[2];1|2} \leq \cdots \leq \sigma^2_{[k];1|2}.$$

It is assumed that the true pairing of the Π_i with the $\sigma^2_{[j];1|2}$ is unknown to the experimenter, and that he has no a priori knowledge which is relevant to the true pairing of any of the populations with the ranked values of the $\sigma^2_{i;1|2}$.

2.1.1 The goal and the probability requirement

In this section we consider the following goal:

(2.2) "To select the Π_i associated with $\sigma^2_{[1];1|2}$."

The term correct selection (CS) will then denote the action of selecting the population associated with $\sigma^2_{[1];1|2}$. (If more than one Π_i is such that the associated $\sigma^2_{i;1|2}$ is equal to $\sigma^2_{[1];1|2}$, then the selection of any one of these Π_i is termed a CS.)

Before experimentation begins the experimenter must specify two constants $\{\theta^*, P^*\}$ with $1 < \theta^* < \infty$ and $1/k < P^* < 1$, which are then incorporated into the probability requirement. The numerical values of these constants are assumed to depend on the economic considerations of the particular problem. The probability requirement can then be stated as

(2.3) PCS $\geq P^*$ whenever $\sigma^2_{[2];1|2} \geq \theta^* \sigma^2_{[1];1|2}$.

The decision procedure which we propose in the next section guarantees the probability requirement when the sample sizes are chosen appropriately.

2.1.2 Single-stage decision procedure

We will base our decision procedure on the values of a predetermined number N of independent observations

$$\underline{x}_i^{(p)} = (x_{i1}^{(p)}, x_{i2}^{(p)}), \qquad (p = 1, 2, \ldots, N) \tag{2.4}$$

on each Π_i $(i = 1, 2, \ldots, k)$, from which we will calculate the values of the k sample conditional variances $s^2_{i;1|2}$, where

$$\bar{x}_{ij} = \sum_{p=1}^{N} x_{ij}^{(p)}/N \qquad (j=1,2), \tag{2.5}$$

$$v_{i;jm} = \sum_{p=1}^{N} (x_{ij}^{(p)} - \bar{x}_{ij})(x_{im}^{(p)} - \bar{x}_{im}) \qquad (j,m=1,2), \tag{2.6}$$

and

$$s^2_{i;1|2} = (v_{i;11} - v^2_{i;12}/v_{i;22})/(N-2). \tag{2.7}$$

We denote the ranked values of the $s^2_{i;1|2}$ by

$$s^2_{[1];1|2} < s^2_{[2];1|2} < \ldots < s^2_{[k];1|2}.$$

In addition we let $s^2_{(i).1|2}$ denote the sample conditional variance associated with $\sigma^2_{[i];1|2}$.

For this single-stage procedure, the experimenter

proceeds as follows: He takes N pairs of observations from each of the k Π_i, computes the values of the $s^2_{i;1|2}$, and selects the population according to the following decision rule:

(2.8) "Select the population associated with $s^2_{[1];1|2}$, and assert that this is the population associated with $\sigma^2_{[1];1|2}$."

2.1.3 Probability of correct selection

The probability of correct selection for this case can be written as

$$(2.9) \qquad PCS = P\{s^2_{(1);1|2} \leq s^2_{(i);1|2} \qquad (i = 2, 3, \ldots, k)\}.$$

By a special case of Anderson's [2] Theorem 4.3.3, we have that each $(N-2)s^2_{(i);1|2}/\sigma^2_{[i];1|2}$ is distributed as chi-square with N-2 degrees of freedom and since these k chi-square variates are independent, this problem reduces to the problem of ranking variances of normal populations already treated by Bechhofer and Sobel [10]. They give exact analytical expressions for the PCS as well as tables for computing the exact sample size when the P* and k values are such that this sample size is small ($N-2 \leq 20$). They also show that when the values of P* and k are such that the sample size is moderately large, a close approximation to the sample size can be calculated from

(2.10) $\sqrt{n} = 2d(k,P^*)/\log \theta^*,$

where $N = n+2$ and $d(k,P^*)$ is given in Table 1 of Bechhofer [5].

2.2 Correlation coefficient formulation

In Section 2.2 we will again consider the problem of selecting the "best" population from a set of k independent bivariate normal populations $\Pi_i (i = 1, 2, \ldots, k)$. However in this section the "goodness" of the Π_i will be measured in terms of the population correlation coefficients ρ_i where, using the notation of Section 2.1,

(2.11) $\rho_i = \sigma_{i;12}/\sqrt{\sigma_{i;11}\,\sigma_{i;22}}.$

We denote the ranked values of the ρ_i by

$$-1 \leq \rho_{[1]} \leq \rho_{[2]} \leq \cdots \leq \rho_{[k]} \leq 1.$$

As before, we assume that the experimenter has no a priori knowledge which is relevant to the true pairing of any of the Π_i with the ranked values of the parameters.

2.2.1 The goal and the probability requirement

For this measure of "goodness" of the Π_i, we consider the following goal:

(2.12) "To select the Π_i associated with $\rho_{[k]}$."

The term correct selection will then denote the action of selecting the population associated with $\rho_{[k]}$. (If more than one Π_i is such that the associated ρ_i is equal to $\rho_{[k]}$, then the selection of any one of these Π_i is termed a CS.)

Before experimentation begins the experimenter must specify two constants $\{\delta^*, P^*\}$ with $0 < \delta^* < 2$ and $1/k < P^* < 1$, which are then incorporated into the probability requirement. For this goal, the probability requirement can then be stated as;

(2.13) PCS $\geq P^*$ whenever $\rho_{[k]} \geq \rho_{[k-1]} + \delta^*$.

The decision procedure which we propose in the next section guarantees the probability requirement when the sample sizes are chosen appropriately.

2.2.2 Single-stage decision procedure

Our single-stage decision procedure is based on the values of the k sample correlation coefficients r_i where, using the notation of Section 2.1.2,

(2.14) $$r_i = v_{i;12} / \sqrt{v_{i;11}\, v_{i;22}}.$$

We denote the ranked values of the r_i by

$$r_{[1]} < r_{[2]} < \cdots < r_{[k]},$$

and in addition we let $r_{(i)}$ denote the sample correlation coefficient associated with $\rho_{[i]}$.

The experimenter proceeds in the same manner as in Section 2.1.2, using the following decision rule in place of (2.8):

(2.15) "Select the population associated with $r_{[k]}$ and assert that this is the population associated with $\rho_{[k]}$."

Eaton [15] has shown that this decision rule is minimax and also is most economical (Hall [31], [32]) within the class of all decision rules.

2.2.3 Probability of correct selection

The PCS for this problem can be written as

$$(2.16) \qquad PCS = P\{r_{(k)} \geq r_{(i)} \qquad (i = 1, 2, \ldots, k-1)\}.$$

Because of the unwieldy form of the exact distribution of the r_i when $\rho_i \neq 0$, we will attack this problem using the asymptotic distribution of Fisher's variance stabilizing transformation

$$(2.17) \qquad z_i = (1/2)\ \log((1 + r_i)/(1 - r_i)).$$

The approach of the distribution of the z_i to normality is much more rapid than that of the distribution of the r_i. The z_i are asymptotically unbiased estimators of the

$$(2.18) \qquad \xi_i = (1/2)\ \log((1 + \rho_i)/(1 - \rho_i)),$$

and have asymptotic variances $1/n$, where $n = N - 3$. (The -3 is a small-sample correction.)

We denote the ranked values of the ξ_i by

$$\xi_{[1]} \leq \xi_{[2]} \leq \cdots \leq \xi_{[k]}.$$

Since ξ_i is a monotonic increasing function of ρ_i, the ranked parameters $\rho_{[i]}$ and $\xi_{[i]}$ are associated with the same population.

In a similar manner we denote the ranked values of the z_i by

$$z_{[1]} < z_{[2]} < \cdots < z_{[k]}.$$

We also let $r_{(i)}$ and $z_{(i)}$ denote the estimators of $\rho_{[i]}$ and $\xi_{[i]}$, respectively, i.e., $\rho_{[i]}$, $\xi_{[i]}$, $r_{(i)}$ and $z_{(i)}$ are all associated with the same population.

Hence, by letting

$$(2.19) \qquad y_i = (\sqrt{n}/2)\,((z_{(i)} - z_{(k)}) - (\xi_{[i]} - \xi_{[k]})),$$

the PCS can be written as

$$(2.20) \qquad PCS = P\{y_i \leq (\sqrt{n}/2)(\xi_{[k]} - \xi_{[i]}) \quad (i=1,2,\ldots,k-1)\}.$$

Since asymptotically ($N \to \infty$) the variates $y_1, y_2, \ldots, y_{k-1}$ have a multivariate normal distribution, the asymptotic probability of correct selection PCS_a can be given as

$$(2.21) \qquad PCS_a = \Phi_{k-1}(\tau_1, \tau_2, \ldots, \tau_{k-1}),$$

where

$$\tau_i = (\sqrt{n}/2)(\xi_{[k]} - \xi_{[i]}),$$

and Φ_{k-1} is a $(k-1)$-variate standard normal distribution function with zero means, unit variances and off-diagonal covariances of 1/2.

The parameter configuration in the region of preference for a correct selection, for which the PCS is minimized is called the least favorable configuration (LFC). Since we will be working with the asymptotic distribution, we will denote the parameter configuration where the PCS_a is minimized by LFC_a.

The following lemma will be used later in proving theorems concerning the LFC_a. Fixing $\rho_{[k]} = \rho'$, we have:

Lemma 2.1

$$\text{(2.22)} \qquad \underset{\rho_{[k-1]} \le \rho' - \delta^*}{\text{infimum}} \quad PCS_a = \Phi_{k-1}(\tau', \tau', \ldots, \tau'),$$

where

$$\tau' = (\sqrt{n}/2)(\xi' - \delta'),$$

$$\xi' = (1/2)\log((1+\rho')/(1-\rho')),$$

and

$$\delta' = (1/2)\log((1+\rho'-\delta^*)/(1-\rho'+\delta^*)).$$

This infimum is attained when

$$\text{(2.23)} \qquad \rho_{[i]} = \rho'-\delta' \qquad (i=1,2,\ldots,k-1).$$

Proof:

We use Rizvi's [45] Theorem 1 and the monotone likelihood ratio property of the normal density to show that PCS_a is a nondecreasing function of $\xi_{[k]}$ and a nonincreasing function of $\xi_{[i]}$ $(i = 1, 2, \ldots, k-1)$ for the decision rule (2.15).

Since $\xi_{[i]}$ is a monotone increasing function of $\rho_{[i]}$, the infimum of the PCS_a is attained at $\xi_{[i]} = \xi' - \delta'$ $(i = 1, 2, \ldots, k-1)$ and we have the desired conclusion.

Using this lemma we now find the LFC_a.

Theorem 2.1

$$\text{(2.24)} \qquad \underset{\rho_{[k]} \geq \rho_{[k-1]} + \delta^*}{\text{infimum}} \quad PCS_a = \Phi_{k-1}(d, d, \ldots, d),$$

where

$$d = (\sqrt{n}/2)\log((1 + \delta^*/2)/(1 - \delta^*/2)).$$

The corresponding LFC_a is

$$\text{(2.25)} \qquad \rho_{[i]} = -\delta^*/2 \qquad (i=1,2,\ldots,k-1)$$

$$\rho_{[k]} = \delta^*/2.$$

Proof:

Using Lemma 2.1, the problem reduces to finding the infimum of τ' when $\delta^*-1 \leq \rho' \leq 1$. Setting the derivative of τ' (taken with respect to ρ') equal to zero we obtain $\rho' = \delta^*/2$.

Differentiating a 2nd time with respect to ρ' and evaluating this expression at $\rho' = \delta^*/2$ we have

$$(1 + \delta^*/2)^{-1}(1 - \delta^*/2)^{-2} - (1 - \delta^*/2)^{-1}(1 + \delta^*/2)^{-2},$$

which can be shown to be > 0 for all $\delta^* > 0$ by noting that

$$(2+\delta^*)/(2-\delta^*) < ((2+\delta^*)/(2-\delta^*))^2.$$

Hence $\rho' = \delta^*/2$ yields a minimum, and the LFC_a is

$$\rho_{[i]} = -\delta^*/2 \qquad (i=1,2,\ldots,k-1)$$

(2.26)

$$\rho_{[k]} = \rho' = \delta^*/2.$$

The asymptotic sample size can be calculated by

$$\sqrt{n} = 2d(k,P^*)/\log\{(1 + \delta^*/2)/(1 - \delta^*/2)\}, \tag{2.27}$$

where $d(k,P^*)$ is given in Table 1 of [5].

2.2.4 An additional restriction

In some cases the experimenter may be interested in guaranteeing the probability requirement only when $\rho_{[k]} \geq \rho^*$, a preassigned constant, or he may have information that $\rho_{[k]} \geq \rho^*$. Formally:

(2.28) PCS $\geq P^*$ whenever $\rho_{[k]} \geq \rho_{[k-1]} + \delta^*$ and $\rho_{[k]} \geq \rho^*$.

For this new probability requirement, it is obvious that when $-1 \leq \rho^* \leq \delta^*/2$, the result of Theorem 3.1 still holds. however, when $\rho^* > \delta^*/2$, we obtain a new LFC_a and a reduction in the asymptotic sample size.

Theorem 2.2

$$\text{infimum} \quad PCS_a = \Phi_{k-1}(d, d, \ldots, d), \tag{2.29}$$

$$\rho_{[k]} \geq \rho_{[k-1]} + \delta^*$$

$$\rho_{[k]} \geq \rho^* > \delta^*/2$$

where

$$d = (\sqrt{n}/2)\log\{(1+\rho^*)(1-\rho^*+\delta^*)/(1-\rho^*)(1+\rho^*-\delta^*)\}.$$

The corresponding LFC_a is

$$\rho_{[i]} = \rho^* - \delta^* \qquad (i=1,2,\ldots,k-1)$$

$$\rho_{[k]} = \rho^* \tag{2.30}$$

Proof:

We again use Lemma 2.1 which reduces the problem to finding the value of ρ' where the infimum PCS_a is attained. Just $\rho^* \leq \rho'$ as before we note that ρ' appears only as an argument of Φ_{k-1} and since Φ_{k-1} is a monotonic increasing function of each of its arguments, we need only minimize τ'. Next we show that the derivative of τ' with respect to ρ' is nonnegative, i.e.,

$$\partial\tau'/\partial\rho' = \{(1+\rho')(1-\rho')\}^{-1} - \{(1+\rho')(1-\rho') + \rho'\delta^* - \delta^{*2}\}^{-1}$$

$$\geq 0,$$

when

$$\delta^*/2 \leq \rho' \leq 1.$$

We use this result along with the result of Theorem 2.1 (unrestricted minimum occurs at $\delta^*/2$) and the continuity of the

function to complete the proof of the theorem.

Again using Table 1 of [5], the asymptotic sample size can be calculated from

$$\sqrt{n} = 2d(k,P^*)/\log\{(1+\rho^*)(1-\rho^*+\delta^*)/(1-\rho^*)(1+\rho^*-\delta^*)\}. \tag{2.31}$$

2.3 Absolute value of the correlation coefficient formulation

In Section 2.3 we consider the absolute value of the population correlation coefficient $\zeta_i = |\rho_i|$ as the measure of "goodness" of the Π_i and are interested in selecting the Π_i with the largest ζ_i from a set of k independent Π_i. We denote the ranked values of the ζ_i by

$$0 \leq \zeta_{[1]} \leq \zeta_{[2]} \leq \cdots \leq \zeta_{[k]} \leq 1.$$

Again, we assume that the experimenter has no a priori knowledge which is relevant to the true pairing of any of the ζ_i with the ranked values of the parameters.

2.3.1 The goal and the probability requirement

The corresponding goal is:

(2.32) "To select the Π_i associated with $\zeta_{[k]}$."

The term correct selection (CS) will then denote the action of selecting the population associated with $\zeta_{[k]}$. (If more than one Π_i is such that the associated ζ_i is equal to $\zeta_{[k]}$, then

the selection of any one of these Π_i is termed a CS.)

The experimenter must specify the same two constants as in Section 2.2.1 (in this case $0 < \delta^* < 1$). The probability requirement can then be stated as

(2.33) $\quad PCS \geq P^*$, whenever $\zeta_{[k]} \geq \zeta_{[k-1]} + \delta^*$.

The decision procedure which we propose in the next section guarantees the probability requirement when the sample sizes are chosen appropriately.

2.3.2 Single-stage decision procedure

Our single-stage decision procedure is based on the absolute values of the k sample correlation coefficients $t_i = |r_i|$, where the r_i are given by (2.14). We denote the ranked values of the t_i, by

$$t_{[1]} \leq t_{[2]} \leq \cdots \leq t_{[k]}.$$

In addition we let $t_{(i)}$ denote the sample quantity associated with $\zeta_{[i]}$.

The experimenter proceeds in the same manner as in Section 2.1.2 using the following decision rule in place of (2.8):

(2.34) "Select the population associated with $t_{[k]}$ and assert that this is the population associated with $\zeta_{[k]}$."

2.3.3 Probability of correct selection

The PCS for this problem can be written as

(2.35) $PCS = P\{t_{(k)} \geq t_{(i)} \qquad (i = 1, 2, \ldots, k-1)\}.$

Because of the unwieldy form of the exact distribution of the t_i when $\rho_i \neq 0$, we will attack this problem using the fact that the asymptotic distribution of z_i (2.17) is normal.

We define

(2.36) $w_i = (1/2)\log((1+t_i)/(1-t_i)),$

(2.37) $\Psi_i = (1/2)\log((1+\zeta_i)/(1-\zeta_i)),$

and denote the ranked values of the w_i and the Ψ_i by

$$w_{[1]} < w_{[2]} < \cdots < w_{[k]}$$

and

$$\Psi_{[1]} \leq \Psi_{[2]} \leq \cdots \leq \Psi_{[k]},$$

respectively. In addition we let the $w_{(i)}$ denote the estimators of the $\Psi_{[i]}$, and hence $t_{(i)}$, $w_{(i)}$, $\zeta_{[i]}$ and $\Psi_{[i]}$ are all associated with the same population.

Since z_i and ξ_i are symmetrical functions of r_i and ρ_i, respectively, about zero, it follows that

(2.38) $w_i = |z_i|$

and

$$\Psi_i = |\xi_i|. \tag{2.39}$$

Using (2.38), (2.39), and some results obtained by Rizvi [45] for ranking the absolute values of means of normal populations, we find the LFC_a for our problem in the following theorem.

Theorem 2.3

The LFC_a which satisfies the conditions of the probability requirement (2.33) is given by

$$\begin{aligned} \zeta_{[i]} &= 0 \qquad (i=1,2,\ldots,k-1) \\ \zeta_{[k]} &= \delta^*. \end{aligned} \tag{2.40}$$

Proof:

Since $\partial\Psi_i/\partial\zeta_i$ and $\partial w_i/\partial t_i$ are positive when $0 \leq \zeta_i \leq 1$ and $0 \leq t_i \leq 1$, respectively, we have by Theorem 1 of Rizvi [45] that the PCS_a is a nonincreasing function of $\zeta_{[k]}$ and a nondecreasing function of $\zeta_{[i]}$ $(i = 1, 2, \ldots, k-1)$. Consequently for any fixed nonnegative value of $\zeta_{[k-1]} = \zeta'$ (say), (2.35) is minimized subject to the restriction of (2.33) by setting

$$\zeta_{[i]} = \zeta' - \delta^* \qquad (i=1,2,\ldots,k-1), \tag{2.41}$$

$$\zeta_{[k]} = \zeta'.$$

The equivalent of Rizvi's Theorem 2 for ζ follows from $\partial\Psi_i/\partial\zeta_i \geq 0$. (Note: Rizvi's θ_i values are our Ψ_i values.) These results along with the LFC given by Rizvi for his problem yield the LFC_a stated in this theorem.

The asymptotic sample size can be calculated from

$$(2.42) \qquad \sqrt{n} = 2\lambda(k,P^*)/\log\{(1+\delta^*)/(1-\delta^*)\},$$

where the λ values are given in Table II of Rizvi [45].

2.3.4 An additional restriction

If we add another restriction to (2.33), similar to that of Section 2.2.5, the new probability requirement is given by

$$(2.43) \qquad PCS \geq P^* \text{ whenever } \zeta_{[k]} \geq \zeta_{[k-1]} + \delta^*$$

$$\text{and } \zeta_{[k]} \geq \zeta^*.$$

Under this new probability requirement, it is obvious that when $0 \leq \zeta^* \leq \delta^*$, the result of Theorem 3.3 still holds. However, when $\zeta^* > \delta^*$, we obtain a new LFC_a with a corresponding reduction in sample size.

Theorem 2.4

The LFC_a which satisfies the conditions of the

probability requirement (2.43) is given by

$$\zeta_{[i]} = \zeta^* - \delta^* \qquad (i=1,2,\ldots,k-1)$$

(2.44)

$$\zeta_{[k]} = \delta^*.$$

Proof:

This proof follows in the same manner as the proof of Theorem 2.3.

The asymptotic sample size can be calculated from

$$\sqrt{n} = 2\lambda(k,P^*)/\log\{(1+\delta^*)(1-\zeta^*+\delta^*)/(1+\zeta^*-\delta^*)(1-\delta^*)\}, \tag{2.45}$$

where the values of λ are given in Table II of [45].

PART II

One (k+1)-Variate Normal Population

In Part II we will consider situations in which we have a random vector $\underline{x} = (x_0, x_1, \ldots, x_k)$ which is a (k+1)-variate normal distribution with unknown mean vector $\underline{\mu} = (\mu_0, \mu_1, \ldots, \mu_k)$ and unknown covariance matrix Σ. Throughout this part we will be interested in predicting the variate x_0 (which we term the predictand) on the basis of the best linear combination of variates in sets of fixed size t of the k ($t < k$) variates $x_1, x_2, \ldots, x_k$ (which we term the predictor variates). For any given set of t predictor variates, the "goodness" of the prediction will be measured in terms of the population conditional variances of x_0 given these t predictor variates, $\sigma^2_{0.i_1,i_2,\ldots,i_t}$ (or equivalently in terms of the population multiple correlation coefficient between x_0 and these t predictor variates, $\bar{R}_{0.i_1,i_2,\ldots,i_t}$). For fixed (k,t) we will be interested in these $U = C^k_t$ parameters.

Ultimately our objective will be to provide a rational basis for determining the sample size for an experiment in which the goal is to select the "best" set of t variates when certain probability requirements (which will be described precisely in Section 3.3) are to be guaranteed.

The general problem of selecting a set of variates to predict a specified variate is an old one. The formulation that we adopt here falls within the framework of the "indifference zone" ranking approach. In this formulation it is necessary to know the exact joint distribution of the statistics on which the decision procedure is based, in order to determine the minimum sample size which guarantees the probability requirement. For all of the problems which we consider in Part II, the determination of these exact distributions appears almost hopeless (and even if one were able to find them, they would be very unwieldy). Thus we will attack these problems by finding the asymptotic joint distribution of these statistics and work with it in the same manner as one would with the exact joint distribution, were it known. That is, we will obtain the infimum of the asymptotic probability of correct selection over the region of preference for a correct selection and choose our sample size in such a way that this infimum satisfies the probability requirement. The asymptotic joint distributions referred to above are derived in Appendix A. We will study the various special cases in the different chapters of Part II. For certain of these cases the exact infimum of the asymptotic probability, referred to above, is also difficult to obtain, and in these cases we will find a lower bound to this asymptotic probability and obtain

the infimum of this lower bound. Using these results, we will find some conservative approximations to the asymptotic sample sizes necessary to satisfy the probability requirements.

3. General notation

In Part II we will consider sets of a preassigned number of t of the k predictor variates and will use the following notation to label these $U = C_t^k$ different sets. Let $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_k)$ be a k-vector consisting of zeros and ones with $\sum_{i=1}^{k} \alpha_i = t$, a given integer $(1 \leq t \leq k-1)$; and let $\underline{x}_\alpha = (x_{i_1}, x_{i_2}, \ldots, x_{i_t})$ be a t-vector obtained from the k-vector $(x_1, x_2, \ldots, x_k)$ of the predictor variates by deleting those x_i for which $\alpha_i = 0$.

In a similar manner we let $\underline{\mu}_\alpha$ and Σ_α denote the population mean vector and the population covariance matrix, respectively, of $\underline{x}_\alpha$ and let $\underline{\sigma}_{0\alpha}$ denote the vector of population covariances between the random variable x_0 and the random vector $\underline{x}_\alpha$.

In addition, we denote the population conditional variance of x_0, given the set of predictor variates $\underline{x}_\alpha$, by

$$\text{Var}(x_0 | \underline{x}_\alpha) = \sigma^2_{0.\alpha} \tag{3.1}$$

and the population multiple correlation coefficient, between x_0 and x_α, by

$$\bar{R}_{0.\alpha} = ((\underline{\sigma}_{0\alpha} \Sigma_\alpha^{-1} \underline{\sigma}'_{0\alpha})/\sigma_{00})^{1/2}, \tag{3.2}$$

so that the conditional variance of x_0, given the set of

predictor variates $\underline{x}_\alpha$, can be written as

$$\sigma^2_{0.\alpha} = \sigma_{00} - \underline{\sigma}_{0\alpha}\Sigma_\alpha^{-1}\underline{\sigma}'_{0\alpha}$$
$$= \sigma_{00}(1 - \bar{R}^2_{0.\alpha}). \tag{3.3}$$

3.1 Ranked values of the parameters--notation

In our formulation of the ranking problem we will be interested in the ranked values of the $\sigma^2_{0.\alpha}$ and of the $\bar{R}_{0.\alpha}$. For fixed t we denote the ranked values of these parameters by

$$\sigma^2_{0.[1]} \leq \sigma^2_{0.[2]} \leq \cdots \leq \sigma^2_{0.[U]}$$

and

$$\bar{R}_{0.[1]} \leq \bar{R}_{0.[2]} \leq \cdots \leq \bar{R}_{0.[U]},$$

where, as before, $U = C^k_t$. Since $\bar{R}_{0.\alpha} \geq 0$, we have

$$\sigma^2_{0.[i]} = \sigma_{00}(1-\bar{R}^2_{0.[U-i+1]}). \tag{3.4}$$

It is assumed that the true pairing of the $\underline{x}_\alpha$ with $\sigma^2_{0.[j]}$ (or equivalently with $\bar{R}_{0.[U-j+1]}$) is unknown to the experimenter and that he has no a priori knowledge which is relevant to the true pairing of any of the populations with the ranked values of the parameters, i.e., it is not known which t of the k variates $x_1, x_2, \ldots, x_k$ are associated with any of the $\sigma^2_{0.[j]}$.

3.2 Goals

In Part II the following two equivalent goals will be of interest to us:

(3.5a) Goal A: "To select the set of t variates associated with $\sigma^2_{0.[1]}$."

(3.5b) Goal B: "To select the set of t variates associated with $\bar{R}_{0.[U]}$."

The term correct selection will denote the action of selecting the set of t variates associated with $\sigma^2_{0.[1]}$ (or equivalently the set of t variates associated with $\bar{R}_{0.[U]}$). (If more than one of the $\sigma^2_{0.\alpha}$ are equal to $\sigma^2_{0.[1]}$, the selection of the set of variates corresponding to any of these $\sigma^2_{0.\alpha}$ is considered a correct selection.)

3.3 Probability requirement

Before experimentation begins the experimenter must specify two constants $\{\theta^*, P^*\}$ with $1 < \theta^* < \infty$ and $1/U < P^* < 1$, which are then incorporated into the probability requirement. The numerical values of these constants are assumed to depend on economic considerations associated with the particular problem. The probability requirement can be stated as

(3.6a) $PCS \geq P^*$ whenever $\sigma^2_{0.[2]}/\sigma^2_{0.[1]} \geq \theta^*$

or equivalently,

(3.6b) PCS $\geq$ P* whenever $(1-\bar{R}^2_{0.[U-1]})/(1-\bar{R}^2_{0.[U]}) \geq \theta^*$.

The decision procedure which we propose in the next section guarantees these probability requirements when the sample sizes are chosen appropriately.

3.4 Ranked values of the statistics--notation

In general we assume that the experimenter will be taking a predetermined number N of independent vector-observations

$$(3.7) \qquad \underline{x}^{(p)} = (x_0^{(p)}, x_1^{(p)}, \ldots, x_k^{(p)}), \qquad (p=1,2,\ldots,N)$$

from a (k+1)-variate normal population. In terms of the N vector-observations, we denote the sample mean of the $i^{\underline{th}}$ variate by

$$(3.8) \qquad \bar{x}_i = \sum_{p=1}^{N} x_i^{(p)}/N,$$

and the vector of sample means and the matrix of sums of squares and cross products of deviations about the mean by

$$(3.9) \qquad \underline{\bar{x}} = (\bar{x}_0, \bar{x}_1, \ldots, \bar{x}_k)',$$

and

$$(3.10) \qquad V = \sum_{p=1}^{N} (\underline{x}^{(p)} - \underline{\bar{x}})'(\underline{x}^{(p)} - \underline{\bar{x}}),$$

respectively.

Using the α notation defined in Section 3, the sample mean vector and the sum of squares and cross products of deviations about the mean matrix of $\underset{\sim}{x}_\alpha$ are given by

$$(3.11) \qquad \bar{\underset{\sim}{x}}_\alpha = \bar{\underset{\sim}{x}}_{i_1, i_2, \ldots, i_t}',$$

and

$$(3.12) \qquad V_\alpha = \sum_{p=1}^{N} (\underset{\sim}{x}_\alpha^{(p)} - \bar{\underset{\sim}{x}}_\alpha)'(\underset{\sim}{x}_\alpha^{(p)} - \bar{\underset{\sim}{x}}_\alpha),$$

respectively. The usual sample quantities associated with σ_{00}, $\underset{\sim}{\sigma}_{0\alpha}^2$, $\sigma_{0.\alpha}^2$, and $\bar{R}_{0.\alpha}$ are given by

$$(3.13) \qquad v_{00} = \sum_{p=1}^{N} (x_0^{(p)} - \bar{x}_0)^2,$$

$$(3.14) \qquad \underset{\sim}{v}_{0\alpha} = \sum_{p=1}^{N} (x_0^{(p)} - \bar{x}_0)(\underset{\sim}{x}_\alpha^{(p)} - \bar{\underset{\sim}{x}}_\alpha),$$

$$(3.15) \qquad s_{0.\alpha}^2 = (v_{00} - \underset{\sim}{v}_{0\alpha} V_\alpha^{-1} \underset{\sim}{v}_{0\alpha}')/(N-t-1),$$

and

$$(3.16) \qquad R_{0.\alpha} = ((\underset{\sim}{v}_{0\alpha} V_\alpha^{-1} \underset{\sim}{v}_{0\alpha}')/v_{00})^{1/2}.$$

We let $n = N - t - 1$ and note that

$$(3.17) \qquad s_{0.\alpha}^2 = v_{00}(1 - R_{0.\alpha}^2)/n.$$

We denote the ranked values of the $s_{0.\alpha}^2$ and the $R_{0.\alpha}$ by

$$s_{0.[1]}^2 < s_{0.[2]}^2 < \ldots < s_{0.[U]}^2$$

and

$$R_{0.[1]} < R_{0.[2]} < \cdots < R_{0.[U]},$$

respectively. Since $R_{0.\alpha} \geq 0$, we have

$$s^2_{0.[i]} = v_{00}(1 - R^2_{0.[U-i+1]})/n. \tag{3.18}$$

In addition we denote the sample conditional variance and the sample multiple correlation coefficient of the (t+1)-variate normal distribution associated with $\sigma^2_{0.[i]}$ and $\bar{R}_{0.[U-i+1]}$ by $s^2_{0.(i)}$ and $R_{0.(U-i+1)}$ respectively, so that

$$s^2_{0.(i)} = v_{00}(1 - R^2_{0.(U-i+1)})/n. \tag{3.19}$$

3.5 Single-stage decision procedure

In Part II we will be concerned with a single-stage procedure which guarantees the probability requirement (3.8). For this single-stage procedure the experimenter proceeds as follows: He takes N independent vector-observations from the (k+1)-variate normal population, computes the values of the U sample conditional variances $s^2_{0.\alpha}$ (or equivalently the U sample multiple correlation coefficients $R_{0.\alpha}$) and selects the set of t variates according to the following decision rule:

(3.20) "Select the set of t variates associated with $s^2_{0.[1]}$ (or equivalently $R_{0.[U]}$) and assert that this set of

variates is associated with $\sigma^2_{0.[1]}$ (or equivalently $\bar{R}_{0.[U]}$)."

3.6 Probability of correct selection

The probability of correct selection (PCS) using the procedure of Section 3.5 can be written as

$$(3.21) \qquad PCS = P\{s^2_{0.(1)} \leq s^2_{0.(i)}; \quad (i = 2, 3, \ldots, U)\},$$

where each of the n $s^2_{0.(i)}/\sigma^2_{0.[i]}$ is distributed as chi-square with n degrees of freedom (Anderson [2], Theorem 4.3.3). However the $s^2_{0.(i)}$ are not independently distributed and their joint distribution does not appear to be known. The exact distribution would appear to be at least as complicated as that of the joint distribution of the simple correlation coefficients in samples from a multivariate normal distribution, which is quite messy. Since knowledge of this distribution is necessary in order to determine the minimum required sample size (see the analogous, but much simpler problem described in Section 2.1.4), we will study this problem from the large-sample point of view. (In practice, "large" samples will usually be required when applying this procedure.) By using the variance stabilizing, logarithmic transformation of the sample conditional variances, we obtain an asymptotic approximation for the PCS which should be

sufficiently accurate to determine the sample size requirement for many problems.

We define

$$(3.22) \qquad y_i' = (\sqrt{n}/2)\log(s^2_{0.(1)}/s^2_{0.(U-i+1)}) \qquad (i=1,2,\ldots,U-1),$$

$$(3.23) \qquad \gamma_i' = -(\sqrt{n}/2)\log(\sigma^2_{0.[1]}/\sigma^2_{0.[U-i+1]}) \qquad (i=1,2,\ldots,U-1),$$

and

$$w_i' = y_i' + \gamma_i' \qquad (i=1,2,\ldots,U-1).$$

Then (3.21) can be written as

$$(3.24) \qquad PCS = P\{w_i' \leq \gamma_i' \qquad (i = 1, 2, \ldots, U-1)\}.$$

Using Theorems A.1 and A.2, we obtain the following asymptotic ($N \to \infty$) approximation for this PCS, which we denote by:

$$(3.25) \qquad PCS_a = \Phi^*_{U-1}(\gamma_1', \gamma_2', \ldots, \gamma_{U-1}'),$$

where Φ^*_{U-1} is a (U-1)-variate normal distribution function having zero means. (The * here is used to indicate that the variances are not unity, i.e., this is not a standardized multivariate normal distribution function.) We have not been able to determine the covariance matrix for general Σ and arbitrary t. However, in Appendix A we give results for general Σ when $t = 1$ and for a special form of Σ (i.e.,

uncorrelated predictor variates) for arbitrary t. These results will be used in the following chapters to determine the asymptotic sample size.

4. Uncorrelated predictor variates

Throughout Chapter 4 we consider the case in which the predictor variates $x_1, x_2, \ldots, x_k$ of the (k+1)-variate normal distribution are uncorrelated, i.e., $\rho_{ij} = 0$, $(i \neq j;\ i = 1, 2, \ldots, k)$. We denote the covariance matrix of this (k+1)-variate normal distribution by Σ_0, where the $ij^{\underline{th}}$ element of Σ_0 is given by

$$(4.1) \qquad (\Sigma_0)_{ij} = \begin{cases} \sigma_{ii}, & (j = i;\ i = 0, 1, \ldots, k) \\ \rho_{ij}\sqrt{\sigma_{ii}\sigma_{jj}}, & (i \neq j;\ i,j = 0, 1, \ldots, k) \end{cases}.$$

The mean vector μ and the nonzero elements of Σ_0 are assumed to be unknown. The assumption of uncorrelated predictor variates yields a simplification in the covariances of the asymptotic joint distribution of the $s^2_{0.i}$ and allows the requirement, that Σ_0 be nonnegative definite, to be expressed in a simple form.

4.1 t = 1

We first consider the situation in which the experimenter's objective is to select t = 1 variate from the k possible predictor variates.

4.1.1 Notation

For this case, the general notation of Chapter 3

simplifies considerably. The k-vector α now consists of k - 1 zeros and 1 one. If we let i denote that component of α which is one, then

(4.2) $\quad \underline{x}_\alpha = x_i,$

(4.3) $\quad \sigma^2_{0.\alpha} = \sigma^2_{0.i},$

(4.4) $\quad \bar{R}_{0.\alpha} = \bar{R}_{0.i},$

and since $\bar{R}^2_{0.\underline{i}} = \rho^2_{0i}$, (3.3) becomes

(4.5) $\quad \sigma^2_{0.i} = \sigma_{00}(1 - \rho^2_{0i}).$

We denote the ranked values of the ρ^2_{0i} by

$$\rho^2_{0[1]} \leq \rho^2_{0[2]} \leq \cdots \leq \rho^2_{0[k]},$$

and write (3.4) as

(4.6) $\quad \sigma^2_{0.[i]} = \sigma_{00}(1 - \rho^2_{0[k-i+1]}).$

Then the goal (3.5) can be stated as:

(4.7) "To select the variate associated with $\rho^2_{0[k]}$,"

and the probability requirement (3.6) becomes

(4.8) $\quad PCS \geq P^*$ whenever $(1 - \rho^2_{0[k-1]})/(1 - \rho^2_{0[k]}) \geq \theta^*.$

Similar notational simplifications result for the

corresponding sample quantities, i.e., (3.12), (3.14), (3.15), and (3.16) become

$$(4.9) \qquad v_{ii} = \sum_{p=1}^{N} (x_i^{(p)} - \bar{x}_0)^2,$$

$$(4.10) \qquad v_{0i} = \sum_{p=1}^{N} (x_0^{(p)} - \bar{x}_0)(x_i^{(p)} - \bar{x}_i),$$

$$(4.11) \qquad s^2_{0.i} = (v_{00} - v_{0i}/v_{ii}^2)/n,$$

and

$$(4.12) \qquad R_{0.i} = \sqrt{v_{0i}^2/v_{00}v_{ii}},$$

respectively. Since $R^2_{0.i} = r^2_{0i}$ and $n = N - 2$, (3.17) becomes

$$(4.13) \qquad s^2_{0.i} = v_{00}(1 - r^2_{0i})/n.$$

Denoting the ranked values of the r^2_{0i} by

$$r^2_{0[1]} < r^2_{0[2]} < \ldots < r^2_{0[k]},$$

(3.18) becomes

$$(4.14) \qquad s^2_{0.[i]} = v_{00}(1 - r^2_{0[k-i+1]})/n.$$

In addition we denote the sample correlation coefficient associated with $\rho^2_{0[i]}$ by $r_{0(i)}$, and hence (3.19) can be written as

$$(4.15) \qquad s^2_{0.(i)} = v_{00}(1 - r^2_{0(k-i+1)})/n.$$

4.1.2 Probability of correct selection

We define

$$\gamma_{ij} = \frac{(1 - \rho^2_{0[i]} - \rho^2_{0[j]})^2}{(1 - \rho^2_{0[i]})(1 - \rho^2_{0[j]})} \qquad (i \neq j; i,j=1,2,\ldots,k), \tag{4.16}$$

$$y_i = \frac{\sqrt{n}\,\log(s^2_{0.(1)}/s^2_{0.(k-i+1)})}{2\sqrt{1 - \gamma_{ik}}} \qquad (i=1,2,\ldots,k-1), \tag{4.17}$$

$$\varepsilon_i = \frac{\sqrt{n}\,\log((1 - \rho^2_{0[i]})/(1 - \rho^2_{0[k]}))}{2\sqrt{1 - \gamma_{ik}}} \qquad (i=1,2,\ldots,k-1), \tag{4.18}$$

and

$$w_i = y_i + \varepsilon_i \qquad (i=1,2,\ldots,k-1).$$

Then the PCS (3.24) can be written as

$$PCS = P\{w_i \leq \varepsilon_i \qquad (i = 1, 2, \ldots, k-1)\}. \tag{4.19}$$

Using Corollary A.4a, we obtain the following asymptotic ($N \to \infty$) approximation for (4.19):

$$PCS_a = \Phi_{k-1}(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{k-1}), \tag{4.20}$$

where Φ_{k-1} is a (k-1)-variate standard normal distribution function with zero means, unit variances, and off-diagonal covariances given by

$$\gamma'_{ij} = \frac{(1 - \gamma_{ik} - \gamma_{jk} + \gamma_{ij})}{2\sqrt{(1 - \gamma_{ik})(1 - \gamma_{jk})}} \qquad (i \neq j; i,j=1,2,\ldots,k-1). \tag{4.21}$$

4.1.3 $k = 2$

In this section we obtain the asymptotic sample size by finding the infimum of the PCS_a over the region of preference for a correct selection for the case ($k = 2$, $t = 1$). The complex nature of the correlation (in the asymptotic distribution) between the estimators $s^2_{0.(1)}$ and $s^2_{0.(2)}$, which depends on the covariance matrix Σ_0 of the original $(k+1)$-variate normal distribution, complicates this problem. For this case (4.20) reduces to

$$(4.22) \qquad PCS_a = \Phi(\varepsilon_1),$$

where Φ $(=\Phi_1)$ is the standard univariate normal distribution function.

Preliminary to finding the infimum of (4.22), we give two lemmas. Lemma 4.1 is a representation of the requirement that Σ_0 be nonnegative definite (n.n.d.) when $k = 2$. Lemma 4.2 shows that ε_i is a decreasing function of $\rho^2_{0[i]}$.

Lemma 4.1

For $k = 2$, Σ_0 n.n.d. is equivalent to the following inequalities:

$$\sigma_{ii} \geq 0 \qquad (i=0,1,2)$$

$$\rho^2_{01} + \rho^2_{02} \leq 1.$$

Proof:

This equivalence is easily established by using the representation that a symmetric matrix is n.n.d. if all of its principal minors are nonnegative.

Lemma 4.2

$$\partial \varepsilon_i / \partial \rho^2_{0[i]} < 0 \qquad (i=1,2,\ldots,k-1), \tag{4.23}$$

when

$$\rho^2_{0[i]} + \rho^2_{0[k]} \leq 1$$

and

$$(1 - \rho^2_{0[i]})/(1 - \rho^2_{0[k]}) \geq \theta^*.$$

Proof:

We let

$$g = (\sqrt{n}/2)\log((1 - \rho^2_{0[i]})/(1 - \rho^2_{0[k]}))$$

and $h = \sqrt{1 - \gamma_{ik}}$, so that $\varepsilon_i = g/h$. Then

$$\partial g/\partial \rho^2_{0[i]} = -\sqrt{n}/(2(1 - \rho^2_{0[i]})) < 0$$

and

$$\partial h/\partial \rho^2_{0[i]} = f/(2h(1 - \rho^2_{0[i]})^2(1 - \rho^2_{0[k]})), \tag{4.24}$$

where

$$f = (1 - 2\rho^2_{0[i]} + \rho^2_{0[i]} - \rho^2_{0[k]}).$$

To show that (4.24) is positive we need only show that $f > 0$. But

$$\partial f/\partial \rho^2_{0[i]} = -2(1 - \rho^2_{0[i]})$$
$$< 0,$$

and hence f is a decreasing function of $\rho^2_{0[i]}$. To show that $f > 0$, we increase $\rho^2_{0[i]}$ until either

$$\rho^2_{0[i]} = (1 - \theta^*) + \theta^*\rho^2_{0[k]}$$

or

$$\rho^2_{0[i]} = 1 - \rho^2_{0[k]},$$

which ever occurs first. In either case the result follows immediately. Combining these results and noting that $h \geq 0$ and $g < 0$, we obtain the desired conclusion.

Theorem 4.1

$$\text{(4.25)} \quad \begin{array}{l} \text{infimum } \Phi(\varepsilon_1) = \Phi(\varepsilon^*), \\ \theta_{12} \geq \theta^* \\ \Sigma_0 \text{ n.n.d.} \end{array}$$

where

$$\varepsilon^* = (\sqrt{n}/2) \log \theta^*$$

and

$$\theta_{12} = (1 - \rho^2_{0[1]})/(1 - \rho^2_{0[2]}).$$

Proof:

For any pair of values of $\rho^2_{0[1]}$ and $\rho^2_{0[2]}$ satisfying the restrictions of the theorem, it is obvious from Lemma 4.2 that $\Phi(\varepsilon_1)$ is a decreasing function of $\rho^2_{0[1]}$. Hence to obtain the infimum of $\Phi(\varepsilon_1)$, we increase $\rho^2_{0[1]}$ until it attains values on one of the two boundaries (Figure 4.1), i.e., on

(4.26a) $$\rho^2_{0[1]} = (1 - \theta^*) + \theta^*\rho^2_{0[2]}$$

or

(4.26b) $$\rho^2_{0[1]} = 1 - \rho^2_{0[2]}.$$

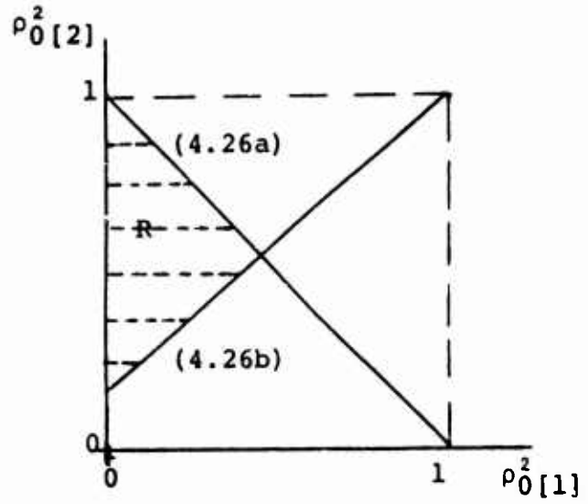
Along either of these boundaries, ε_1 can be expressed as a function of just one of the pair of parameters $\rho^2_{0[1]}$ and $\rho^2_{0[2]}$.

Case (a)

After solving (4.26a) for $\rho^2_{0[2]}$, and substituting this expression into (4.16), we obtain expressions for γ_{12} and ε_1

Figure 4.1

Preference Region (R) for a Correct Selection



on boundary (a) which we denote by γ_a and ε_a, respectively, i.e.,

$$\gamma_a = \frac{(\theta^*(1 - \rho^2_{0[1]}) - \theta^{*2}\rho^2_{0[1]})^2}{(1 - \rho^2_{0[i]})^2}$$

and

$$\varepsilon_a = (\sqrt{n} \log \theta^*)/(2\sqrt{1 - \gamma_a}).$$

Since $\log \theta^* > 0$, $(1 - \gamma_a) \geq 0$ and

$$\partial\gamma_a/\partial\rho^2_{0[1]} = -2\theta^*/(1 - \rho^2_{0[1]})$$
$$< 0,$$

we have

$$\partial \varepsilon_a / \partial \rho^2_{0[1]} = \frac{(\sqrt{n} \log \theta^*) \partial \gamma_a / \partial \rho^2_{0[1]}}{2\sqrt{1 - \gamma_a}}$$
$$< 0.$$

Hence ε_a is a decreasing function of $\rho^2_{0[1]}$ and the infimum of ε_a, along boundary (a) and subject to the conditions of the theorem, occurs at the intersection of boundary (a) and boundary (b).

Case (b)

In a similar manner along boundary (b), after solving (4.26b) for $\rho^2_{0[2]}$ and substituting into (4.16) and (4.18), we obtain $\gamma_b = 0$ and

$$\varepsilon_b = (\sqrt{n}/2) \log((1 - \rho^2_{0[1]})/\rho^2_{0[1]}).$$

Since

$$\partial \varepsilon_b / \partial \rho^2_{0[1]} = -\sqrt{n}/(2\rho^2_{0[1]}(1 - \rho^2_{0[1]}))$$
$$< 0,$$

ε_b is a decreasing function of $\rho^2_{0[1]}$; and the infimum of ε_b, along boundary (b) and subject to the conditions of the theorem, also occurs at the intersection of boundary (a) and boundary (b).

Hence the infimum of $\Phi(\varepsilon_1)$ over this region is attained at the intersection of the two boundaries. Solving (4.26a)

and (4.26b) for $\rho^2_{0[1]}$ and $\rho^2_{0[2]}$, we obtain the LFC (Section 2.2.3)

$$\rho^2_{0[1]} = 1/(1 + \theta^*)$$
$$\rho^2_{0[2]} = \theta^*/(1 + \theta^*), \tag{4.27}$$

and the conclusion of the theorem follows.

Using the result of this theorem, an asymptotic approximation to the sample size can be obtained by setting the r.h.s. of (4.25) equal to P*, from which we obtain

$$\sqrt{n} = 2\Phi^{-1}(P^*)/\log \theta^*, \tag{4.28}$$

where Φ^{-1} is the inverse of the standard normal distribution function Φ.

Alternate Proof:

The theorem can also be proven indirectly by noting that $\gamma_{12} \geq 0$, whenever the restrictions of the theorem are satisfied and hence

$$\text{infimum} \quad \Phi(\varepsilon_1) \geq \Phi(\varepsilon^*).$$
$$\theta_{12} \geq \theta^* \tag{4.29}$$
$$\Sigma_0 \text{ n.n.d.}$$

Substituting the parameter values given by (4.27) into (4.16) and (4.18), we see that the equality is attained and the proof

is complete.

The rather long direct proof was given because it provides some insight into the solution of more general problems considered in later sections. In addition some of the lemmas associated with this theorem are used in the proofs of some of the theorems concerning these problems.

4.1.4 $\underline{k > 2}$

For $k > 2$ the problem of finding the infimum of the PCS_a (4.20) is complicated by the fact that both the ε_i and the γ_i, are functions of the $\rho^2_{0[i]}$.

To illustrate these complications, we briefly consider the case $k = 3$. Using the results of Plackett [43] for the reduction of multivariate normal integrals, we have†

$$(4.30) \qquad PCS_a = \Phi_2(\varepsilon_1, \varepsilon_2)$$

$$= \Phi(\varepsilon_1)\Phi(\varepsilon_2) + H(\rho^2_{0[1]}, \rho^2_{0[2]}),$$

where

$$(4.31) \qquad H(\rho^2_{0[1]}, \rho^2_{0[2]}) = (1/2\pi) \int_0^{\gamma'_{12}} (1/\sqrt{1 - \lambda^2})$$

---

†Gnanadesikan [23] has used this method to obtain some numerical results for the "subset" approach to some selection procedure problems involving a multivariate normal distribution.

$$\exp \frac{-(\varepsilon_1^2 + \varepsilon_2^2 - 2\lambda\varepsilon_1\varepsilon_2)}{2(1-\lambda^2)} \, d\lambda.$$

The difficulties encountered in finding the infimum of the PCS_a using this expression become apparent by taking the derivatives of (4.30) with respect to the $\rho_{0[1]}^2$. However we do note that by combining Lemma 4.5 and (4.30) we obtain the following inequality:

$$(4.32) \qquad \Phi_2(\varepsilon_1, \varepsilon_2) \geq \Phi(\varepsilon_1)\Phi(\varepsilon_2),$$

which is a special case of Theorem 4.3.

Because of the complexity of the PCS_a expression for $k = 3$, we circumvent this problem for $k > 2$ by finding a lower bound for the PCS_a. Preliminary to this we prove three lemmas from which this bound will follow directly as Theorem 4.2.

Lemma 4.3

Σ_0 n.n.d. is equivalent to the following inequalities:

$$\sigma_{ii} \geq 0 \qquad (i=0,1,\ldots,k)$$

$$\sum_{i=1}^{k} \rho_{0i}^2 \leq 1.$$

Proof:

We denote the submatrix of Σ_0 formed by deleting the p $(1 \leq p < k)$ rows $i_1, i_2, \ldots, i_p$ and columns

$j_1, j_2, \ldots, j_p$ by

$$\Sigma_0^{i_1, i_2, \ldots, i_p; j_1, j_2, \ldots, j_p}$$

$$0 \leq i_1 < i_2 < \ldots < i_p \leq k$$

$$0 \leq j_1 < j_2 < \ldots < j_p \leq k,$$

and its determinant by

$$|\Sigma_0^{i_1, i_2, \ldots, i_p; j_1, j_2, \ldots, j_p}|.$$

Using this notation the determinant of Σ_0 is given by

$$|\Sigma_0| = \sigma_{kk} |\Sigma_0^{k;k}| + (-1)^{k-1} \rho_{0k} \sqrt{\sigma_{00}\sigma_{kk}} |\Sigma_0^{0;k}|.$$

By assumption

$$|\Sigma_0^{k;k}| = (1 - \sum_{i=1}^{k-1} \rho_{0i}^2) \prod_{i=0}^{k-1} \sigma_{ii}.$$

$|\Sigma_0^{0;k}|$ and the appropriate lower order minors can be evaluated in the same manner as $|\Sigma_0|$. Continuing in this fashion we obtain

$$|\Sigma_0| = (1 - \sum_{i=1}^{k} \rho_{0i}^2) \prod_{i=0}^{k} \sigma_{ii}.$$

Using this same method of evaluation, we also note that any $(k+1-r)^{\underline{th}}$ $(1 \leq r \leq k)$ principal minor of Σ_0 (a minor formed by deleting the same r rows and columns) can be of two possible forms.

Case (a)

If the $0^{\underline{th}}$ row and column are deleted, the principal minor is given by

$$\prod_{i=1}^{k} \sigma_{ii}.$$

$$i \neq i_1, i_2, \ldots, i_r$$

Case (b)

If the $0^{\underline{th}}$ row and column are not deleted, the submatrix is of the same form as Σ_0 and the corresponding principal minor is given by

$$(1 - \sum_{i=1}^{k} \rho_{0i}^2) \prod_{i=1}^{k} \sigma_{ii}.$$

$$(i \neq i_1, i_2, \ldots, i_r)$$

The proof can be completed by using the representation that symmetric matrices are n.n.d. iff all principal minors are nonnegative and noting the expressions for $|\Sigma_0|$ and the principal minors of Σ_0.

Lemma 4.4

(4.33) $\gamma_{ij} \geq \gamma_{ik}$ $(i < j < k)$,

when

(4.34) $\rho^2_{0[i]} + \rho^2_{0[j]} + \rho^2_{0[k]} \leq 1.$

Proof:

Let

$$f = \gamma_{ij} - \gamma_{ik}.$$

Then after canceling terms, we have

(4.35)
$$f = \rho^4_{0[j]} - \rho^2_{0[j]} - \rho^4_{0[k]} + \rho^2_{0[k]} - \rho^4_{0[i]}\rho^2_{0[k]} - \rho^4_{0[j]}\rho^2_{0[k]} + \rho^4_{0[i]}\rho^2_{0[j]} + \rho^4_{0[k]}\rho^2_{0[j]}.$$

We note that

$$\begin{aligned} \partial f/\partial \rho^2_{0[i]} &= -2\rho^2_{0[i]}\rho^2_{0[k]} + 2\rho^2_{0[i]}\rho^2_{0[j]} \\ &= 2\rho^2_{0[i]}(\rho^2_{0[j]} - \rho^2_{0[k]}) \\ &\leq 0. \end{aligned}$$

Hence f is a nonincreasing function of $\rho^2_{0[i]}$. To show that $f \geq 0$ we increase $\rho^2_{0[i]}$ until either

(4.36a) $$\rho^2_{0[i]} = 1 - \rho^2_{0[j]} - \rho^2_{0[k]}$$

or

(4.36b) $$\rho^2_{0[i]} = \rho^2_{0[j]},$$

whichever occurs first.

Case (a)

Substituting (4.36a) into (4.35) and denoting this expression by f_a, we have, after canceling terms

$$f_a = (\rho_{0[k]}^4 - \rho_{0[j]}^4) - (\rho_{0[k]}^6 - \rho_{0[j]}^6)$$

$$\geq (\rho_{0[k]}^4 - \rho_{0[j]}^4) - \rho_{0[j]}^2(\rho_{0[k]}^4 - \rho_{0[j]}^4)$$

$$\geq 0.$$

Case b

Substituting (4.36b) into (4.35) and denoting this expression by f_b, we have

$$f_b = \rho_{0[j]}^6 + \rho_{0[j]}^4 - \rho_{0[j]}^2 - \rho_{0[k]}^4 + \rho_{0[k]}^2 - 2\rho_{0[j]}^4\rho_{0[k]}^2 + \rho_{0[k]}^4\rho_{0[j]}^2, \tag{4.37}$$

and

$$\partial f_b/\partial \rho_{0[j]}^2 = 3\rho_{0[j]}^4 + 2\rho_{0[j]}^2 - 1 - 4\rho_{C[j]}^2\rho_{0[k]}^2 + \rho_{0[k]}^4.$$

It is evident that

$$\partial f_b/\partial \rho_{0[j]}^2 \leq 0,$$

when $\rho_{0[k]}^2 \leq 1/4$. From (4.34) and (4.36b) we obtain

$$\rho_{0[j]}^{8} \leq (1 - \rho_{0[k]}^{8})/2$$

or

$$\rho_{0[k]}^{2} \leq 1 - 2\rho_{0[j]}^{2},$$

so that

$$\partial f_b/\partial \rho_{0[j]}^{2} \leq 7\rho_{0[j]}^{4} - 3\rho_{0[j]}^{2},$$

when $\rho_{0[k]}^{2} \geq 1/4$. But (4.34) and (4.36b) imply that $\rho_{0[j]}^{2} \leq 3/8$ and hence

$$\partial f_b/\partial \rho_{0[j]}^{2} \leq 0.$$

Using the same method as before, we increase $\rho_{0[j]}^{2}$ until

(4.38b') $\rho_{0[j]}^{2} = \rho_{0[k]}^{2}$

or

(4.38b") $\rho_{0[j]}^{2} = (1 - \rho_{0[k]}^{2})/2,$

whichever occurs first.

Case b'

Substituting (4.38b') into (4.37) and denoting this expression by $f_{b'}$, we obtain $f_{b'} = 0$.

Case b"

Substituting (4.38b") into (4.37) and denoting this expression by $f_{b"}$, we obtain

$$f_{b"} = (-9\rho_{0[k]}^6 + 9\rho_{0[k]}^4 + \rho_{0[k]}^2 - 1)/8.$$

For this case (4.38b") also implies

$$1/3 \leq \rho_{0[k]}^2 \leq 1,$$

which gives us $f_{b"} \geq 0$ and completes the proof.

Lemma 4.5

(4.39) $\gamma'_{ij} \geq 0 \qquad (i < j < k),$

when

$$\rho_{0[i]}^2 + \rho_{0[j]}^2 + \rho_{0[k]}^2 \leq 1.$$

Proof:

The proof follows directly from Lemma 4.4 by noting that $\gamma_{jk} \leq 1$. (γ_{jk} is the covariance between two random variables each having unit variance.) Hence

$$1 - \gamma_{ik} - \gamma_{jk} + \gamma_{ij} \geq 1 - \gamma_{ik}$$

$$\geq 0,$$

and the lemma follows immediately.

Theorem 4.2

$$\Phi_{k-1}(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{k-1}) \geq \prod_{i=1}^{k-1} \Phi(\varepsilon_i) \tag{4.40}$$

Proof:

The proof follows directly from Lemma 4.3 and Slepian's inequality (Lemma B.2).

We use this theorem to find a lower limit for the PCS_a over the region of preference for a correct selection and thus obtain a conservative approximation to the asymptotic sample size. (This approximation is conservative in that it will always be greater than the true asymptotic sample size.)

Theorem 4.3

$$\begin{aligned} &\text{infimum} \quad PCS_a > \{\Phi(\varepsilon^*)\}^{k-1}, \\ &\theta_{k-1,k} \geq \theta^* \\ &\Sigma_0 \text{ n.n.d.} \end{aligned} \tag{4.41}$$

where

$$\theta_{i,k} = (1 - \rho^2_{0[i]})/(1 - \rho^2_{0[k]}) \qquad (i=1,2,\ldots,k-1)$$

and

$$\varepsilon^* = (\sqrt{n}/2) \log \theta^*.$$

Proof:

This theorem follows from Theorem 4.2 and the fact that $\gamma_{ik} \geq 0$ $(i = 1, 2, \ldots, k-1)$.

It is disconcerting to note, however, that

$$\text{infimum} \prod_{i=1}^{k-1} \Phi(\varepsilon_i) > \{\Phi(\varepsilon^*)\}^{k-1}$$

$$\theta_{k-1,k} \geq \theta^* \tag{4.42}$$

$$\Sigma_0 \text{ n.n.d.}$$

for general k.

Setting the r.h.s. of (4.42) equal to P* we obtain a conservative asymptotic approximation to the sample size

$$\sqrt{n} = 2\Phi^{-1}(P^{*1/(k-1)})/\log \theta^*. \tag{4.43}$$

If the experimenter has a priori knowledge that

$$\rho^2_{0[k]} \leq \rho^{*2}, \tag{4.44}$$

where ρ^{*2} is a preassigned constant satisfying

$$\rho^{*2} \leq \{(k-1)\theta^* - (k-2)\}/\{(k-1)\theta^* + 1\}, \tag{4.45}$$

a stronger result can be obtained.

Theorem 4.4

$$\text{(4.46)}\qquad \begin{aligned} &\text{infimum } PCS_a = \{\Phi(\varepsilon^{**})\}^{k-1}, \\ &\theta_{k-1,k} \geq \theta^* \\ &\Sigma_0 \text{ n.n.d.} \\ &\rho^2_{0[k]} \leq \rho^{*2} \end{aligned}$$

where

$$\varepsilon^{**} = \varepsilon^*/\sqrt{1 - \gamma^*}$$

$$= (\sqrt{n} \log \theta^*)/(2\sqrt{1 - \gamma^*}),$$

$$\gamma^* = \{\theta^* - \rho^{*2}(1 - \theta^*)\}/\{\theta^*(1 - \rho^{*2})(1 + \rho^{*2})\},$$

and $\theta_{k,k-1}$ is given in Theorem 4.3.

Proof:

The proof follows directly from Lemma 4.2 by noting that $\rho^2_{0[k]} = \rho^{*2}$ and $\theta_{ik} = \theta^*$ (i = 1, 2, ..., k-1) yield the r.h.s. of (4.46) and Σ_0 is n.n.d.

Setting the r.h.s. of (4.46) equal to P*, we obtain the following approximation to the asymptotic sample size using the a priori information (4.44):

$$\text{(4.47)}\qquad \sqrt{n} = 2\Phi^{-1}(P^{*1/(k-1)})/(\sqrt{1 - \gamma^*} \log \theta^*).$$

4.2 ($k > 2$, $t > 1$)

In the remaining sections of Chapter 4 we consider the problem of selecting a set of predictor variates of fixed but arbitrary number t ($1 < t < k$). We have not been able to show that the conditions required for the Slepian inequality (Lemma B.2) are satisfied (although we conjecture that they do hold). Hence we use the Bonferroni inequality (Lemma B.1) to obtain a lower bound on the PCS_a and proceed in a manner similar to that of Section 4.1.4.

4.2.1 Notation

Using the α notation of Chapter 3, we define the k-vector α^* in terms of the two k-vectors α and α'' by giving the i^{th} component of α^* as

$$\alpha_i^* = \max(\alpha_i, \alpha_i'') . \tag{4.48}$$

We let u_α denote the ordered p-tuple ($t \leq p < \min(k, 2t)$) whose i^{th} component is given by the position number of the i^{th} nonzero component of the vector α. For example, if ($k = 5$, $t = 3$), $\alpha = (1, 0, 0, 1, 1)$, and $\alpha'' = (1, 0, 1, 0, 1)$, then $\alpha^* = (1, 0, 1, 1, 1)$, $u_\alpha = (1, 4, 5)$, $u_{\alpha''} = (1, 3, 5)$, and $u_{\alpha^*} = (1, 3, 4, 5)$.

Since the predictor variates are uncorrelated, we then have

$$(4.49) \qquad \bar{R}^2_{0.\alpha} = \sum_{j \varepsilon u_\alpha} \rho^2_{0j},$$

where $j \varepsilon u_\alpha$ means that j takes on the values of the components of u_α. Hence (3.3) becomes

$$(4.50) \qquad \sigma^2_{0.\alpha} = \sigma_{00}(1 - \sum_{j \varepsilon u_\alpha} \rho^2_{0j}).$$

The population multiple correlation coefficient between x_0 and the set consisting of those predictor variates associated with $\bar{R}_{0.\alpha}$ or $\bar{R}_{0.\alpha''}$ is denoted by $\bar{R}_{0.\alpha^*}$ and can be written as

$$(4.51) \qquad \bar{R}_{0.\alpha^*} = \sum_{j \varepsilon u_{\alpha^*}} \rho^2_{0j}.$$

For the previous example we then have

$$\bar{R}^2_{0.\alpha} = \rho^2_{01} + \rho^2_{04} + \rho^2_{05}$$

$$(4.52) \qquad \bar{R}^2_{0.\alpha''} = \rho^2_{01} + \rho^2_{03} + \rho^2_{05}$$

$$\bar{R}^2_{0.\alpha^*} = \rho^2_{01} + \rho^2_{03} + \rho^2_{04} + \rho^2_{05}.$$

We denote the ranked values of the $\sigma_{0.\alpha}$ and $\bar{R}_{0.\alpha}$, and their estimators in the usual manner. And we let $\sum_{m}^{i,j} \rho^2_{0m}$ denote the summation of the ρ^2_{0m} over the values of the subscript of the variates contained in the union of the two sets of predictor variates associated with $\bar{R}_{0.[i]}$ and $\bar{R}_{0.[j]}$. Thus, if ($k = 5$, $t = 3$) and the variates associated with $\bar{R}_{0.[i]}$ and

$\bar{R}_{0.[j]}$ are x_1, x_2, x_4 and x_1, x_3. x_4, respectively, then

$$(4.53) \qquad \sum_{m}^{i,j} \rho_{0m}^2 = \sum_{m=1}^{4} \rho_{0m}^2$$

4.2.2 Probability of correct selection

Using the notation of Section 4.2.1, we define

$$(4.54) \qquad \lambda_{ij} = \frac{1 - \sum_{m}^{i,j} \rho_{0m}^2}{(1 - \bar{R}_{0.[i]}^2)(1 - \bar{R}_{0.[j]}^2)} \quad (i \neq j; i,j = 1,2,\ldots,U-1),$$

$$(4.55) \qquad y_i = \frac{\sqrt{n}\, \log(s_{0.(1)}^2 / s_{0.(U-i+1)}^2)}{2\sqrt{1 - \lambda_{iU}}} \qquad (i = 1,2,\ldots,U-1),$$

$$(4.56) \qquad \xi_i = \frac{\sqrt{n}\, \log((1 - \bar{R}_{0.[i]}^2)/(1 - \bar{R}_{0.[U]}^2))}{2\sqrt{1 - \lambda_{iU}}}$$

$$(i = 1,2,\ldots,U-1),$$

and

$$w_i = y_i + \xi_i \qquad (i = 1,2,\ldots,U-1).$$

Then (3.24) can be written as

$$(4.57) \qquad PCS = P\{w_i \leq \xi_i \qquad (i = 1, 2, \ldots, U-1)\}$$

Using Corollary A.8, we obtain the following asymptotic ($N \to \infty$) approximation for (4.57).

$$(4.58) \qquad PCS_a = \Phi_{U-1}(\xi_1, \xi_2, \ldots, \xi_{U-1}),$$

where Φ_{U-1} is a standard (U-1)-variate normal distribution function having zero means, unit variances and off-diagonal covariances λ'_{ij} given by

$$(4.59) \qquad \lambda'_{ij} = \frac{(1 - \lambda_{iU} - \lambda_{jU} + \lambda_{ij})}{2\sqrt{(1 - \lambda_{iU})(1 - \lambda_{jU})}} \qquad (i \neq j; i,j = 1,2,\ldots,U-1)$$

Using this expression for the PCS_a, we obtain the following theorem.

Theorem 4.5

$$(4.60) \qquad \begin{aligned} &\text{infimum} \quad PCS_a \geq 1 - (U-1)\Phi(-\xi^*), \\ &\theta_{U-1,U} \geq \theta^* \\ &\Sigma_0 \text{ n.n.d.} \end{aligned}$$

where

$$\xi^* = (\sqrt{n}/2) \log \theta^*$$

and

$$\theta_{U-1,U} = (1 - \bar{R}^2_{0.[U-1]})/(1 - \bar{R}^2_{0.[U]}).$$

Proof:

The proof follows from the fact that $\lambda_{iU} \geq 0$ $(i = 1, 2, \ldots, U-1)$.

Setting the r.h.s. of (4.60) equal to P*, we obtain an approximation to the asymptotic sample size

$$\sqrt{n} = -2\Phi^{-1}((1 - P^*)/(U-1))/\log \theta^*. \tag{4.61}$$

Although there is no general set of parameter configurations where the lower bound attains this lower limit for all t and we do not wish to analyze each t separately, we do note that for the case t = k-1 this lower limit is attained. For this case we have U = k and using Corollary A.8, we obtain the following theorem.

Theorem 4.6

$$\begin{aligned} &\text{infimum} \quad 1 - \sum_{i=1}^{k-1} \Phi(\xi_i) = 1 - (k-1)\Phi(-\xi^*), \\ &\theta_{k,k-1} \geq \theta^* \\ &\Sigma_0 \text{ n.n.d.} \end{aligned} \tag{4.62}$$

where

$$\xi^* = (\sqrt{n}/2) \log \theta^*$$

and

$$\theta_{k-1,k} = (1 - \bar{R}^2_{0.[k-1]})/(1 - \bar{R}^2_{0.[k]}).$$

Proof:

Since $\lambda_{ik} \geq 0$ (i = 1, 2, ..., k-1) implies that Theorem

4.6 holds when the = in (4.62) is replaced by $\geq$, we need only exhibit a parameter configuration for which the lower limit is attained. Since t = k-1, we have

$$\sum_{m}^{i,j} \rho_{0m}^2 = \sum_{i=1}^{k} \rho_{0[i]}^2 .$$

Hence for the parameter configuration

$$\begin{aligned} \rho_{0[1]}^2 &= 1/(1 + (k-1)/\theta^*) \\ \rho_{0[i]}^2 &= \theta^* \rho_{0[1]}^2 \qquad (i=2,3,\ldots,k-1), \end{aligned} \tag{4.63}$$

we have $\lambda_{ik} = 0$ $(i = 1, 2, \ldots, k-1)$ and

$$\xi_i = (\sqrt{n}/2) \log \theta^* \qquad (i=1,2,\ldots,k-1),$$

which completes the proof of the theorem.

5. Predictor variates with unknown correlations

In this chapter we consider the case where the correlation structure of the predictor variates of the (k+1)-variate normal distribution is also unknown. The Σ n.n.d. requirement does not reduce to a simple set of inequalities for arbitrary k as did the Σ_0 n.n.d. requirement, although the additional parameters ρ_{ij} ($i \neq j$; $i,j = 1, 2, \ldots, k$) make the Σ n.n.d. requirement less stringent.

We consider only the case of $t = 1$. For $t > 1$, the correlation between the sample conditional variances becomes extremely messy. Throughout this chapter, we use the notation given in Section 4.1.1.

5.1 Probability of correct selection

For $t = 1$, we have $n = N-2$. We define $\rho_{[i][j]}$ to be the correlation between the predictor variates associated with $\rho^2_{0[i]}$ and $\rho^2_{0[j]}$, respectively and also define

$$(5.1) \qquad \omega_{ij} = \frac{(1 - \rho^2_{0[i]} - \rho^2_{0[j]} + \rho_{0[i]}\rho_{0[j]}\rho_{[i][j]})^2}{(1 - \rho^2_{0[i]})(1 - \rho^2_{0[j]})} \qquad (i \neq j; i,j = 1,2,\ldots,k),$$

$$(5.2) \qquad y_i = \frac{\sqrt{n}\,\log(s^2_{0.(1)}/s^2_{0.(k-i+1)})}{2\sqrt{1-\omega_{ik}}} \qquad (i=1,2,\ldots,k-1),$$

$$(5.3) \qquad \tau_i = \frac{\sqrt{n}\,\log((1-\rho^2_{0[i]})/(1-\rho^2_{0[k]}))}{2\sqrt{1-\omega_{ik}}} \qquad (i=1,2,\ldots,k-1),$$

and $$w_i = y_i + \tau_i \qquad (i=1,2,\ldots,k-1).$$

Then the PCS (3.24) can be written as

$$(5.4) \qquad PCS = P\{w_i \leq \tau_i \qquad (i = 1, 2, \ldots, k-1)\}.$$

Using Corollary A.4 and (5.4), we can write the PCS_a (3.25) as

$$(5.5) \qquad PCS_a = \Phi_{k-1}(\tau_1, \tau_2, \ldots, \tau_{k-1}),$$

where Φ_{k-1} is the (k-1)-variate standard normal distribution function with zero means, unit variances, and off-diagonal covariances given by

$$(5.6) \qquad \omega'_{ij} = \frac{(1-\omega_{ik}-\omega_{jk}+\omega_{ij})}{2\sqrt{(1-\omega_{ik})(1-\omega_{jk})}} \qquad (i\neq j; i,j=1,2,\ldots,k-1).$$

5.2 k = 2

We first consider the case (k = 2, t = 1), where the PCS_a (5.5) reduces to

(5.7) $\quad PCS_a = \Phi(\tau_1).$

Preliminary to finding the infimum of this expression we prove a lemma which gives a simple representation of the Σ n.n.d. requirement for $k = 2$.

Lemma 5.1

For $k = 2$ the requirement that Σ be n.n.d. is equivalent to the following inequalities:

(5.8)
$$\sigma_{ii} \geq 0 \qquad (i=0,1,2)$$
$$\rho^2_{0[1]} + \rho^2_{0[2]} + \rho^2_{[1][2]} - 2\rho_{0[1]}\rho_{0[2]}\rho_{[1][2]} \leq 1.$$

Proof:

As before, this equivalence can be established by using the representation that a symmetric matrix is n.n.d. iff all principal minors are nonnegative.

Theorem 5.1

(5.9)
$$\text{infimum} \quad PCS_a = \Phi(\tau^*),$$
$$\theta_{12} \geq \theta^*$$
$$\Sigma \text{ n.n.d.}$$

where

$$\tau^* = (\sqrt{n}/2) \log \theta^*$$

and

$$\theta_{i2} = (1 - \rho^2_{0[1]})/(1 - \rho^2_{0[2]})$$

Proof:

Since $\omega_{12} \geq 0$, the theorem holds when the = of (5.9) is replaced by $\geq$. Hence we need only exhibit a parameter configuration satisfying the two restrictions of the theorem and for which $\omega_{12} = 0$. One such parameter configuration is given in (4.27).

The asymptotic sample size is the same as given by (4.28). It is also interesting to note that for any given θ^*, the infimum can be attained for more than one parameter configuration, whereas in Theorem 4.1, for any given θ^*, the infimum was attained for only one parameter configuration.

5.3 $k > 2$

We have been unable to show that the conditions (Lemma B.2) required for the use of the Slepian inequality hold when the predictor variates are no longer assumed to be uncorrelated. Hence we use the Bonferroni inequality to obtain a lower bound on the PCS_a. Using (5.5) and (B.1) we have

$$PCS_a \geq 1 - \sum_{i=1}^{k-1} \Phi(-\tau_i). \tag{5.10}$$

In the following theorem we find the infimum of this

lower bound and use this result to obtain the corresponding approximation to the asymptotic sample size.

Theorem 5.2

$$\text{infimum} \quad \left(1 - \sum_{i=1}^{k-1} \Phi(\tau_i)\right) = 1 - (k-1)\Phi(-\tau^*)$$

$$\theta_{k-1,k} \geq \theta^* \qquad (5.11)$$

$$\Sigma \text{ n.n.d.}$$

where

$$\tau^* = (\sqrt{n}/2) \log \theta^*$$

and

$$\theta_{k-1,k} = (1 - \rho^2_{0[k-1]})/(1 - \rho^2_{0[k]}).$$

Proof:

Since $\omega_{ik} \geq 0$ $(i = 1, 2, \ldots, k)$, we note that the theorem holds when the = of (5.11) is replaced by $\geq$. Hence we need only exhibit a parameter configuration satisfying the two restrictions of the theorem and for which $\omega_{ik} = 0$ $(i = 1, 2, \ldots, k-1)$. One such parameter configuration is given by

$$\rho_{0[i]} = \sqrt{1/(1+\theta^*)} \qquad (i=1,2,\ldots,k-1)$$

$$\rho_{0[k]} = \sqrt{\theta^*/(1+\theta^*)}$$

(5.12)

$$\rho_{[i][j]} = 1 \qquad (i\neq j; i,j=1,2,\ldots,k-1)$$

$$\rho_{[i][k]} = 0 \qquad (i=1,2,\ldots,k-1).$$

The restriction $\theta_{k-1,k} \geq \theta^*$ and $\omega_{ik} = 0$ ($i = 1, 2, \ldots, k-1$) follow directly. To prove the existence of a nonnegative definite covariance matrix corresponding to this configuration, we assume without loss of generality that the predictor variates are ordered according to increasing ρ_{0i}^2, so that x_i is the variate associated with $\rho_{0[i]}^2$.

We use the result (e.g., Anderson [2]) that a $k+1$ by $k+1$ symmetric matrix Σ is n.n.d. if there exists a $k+1$ by r matrix A, where $r < k+1$ is the rank of Σ, such that $T = A\Sigma A'$ and T is positive definite.

For our problem, $r = 2$ and we have

$$(5.13) \qquad A = \begin{vmatrix} 1\ 0\ 0\ 0\ \ldots,\ 0 \\ 0\ 1\ 0\ 0\ \ldots,\ 0 \end{vmatrix},$$

so that

$$(5.14) \qquad T = \begin{vmatrix} 1 & \sqrt{1/(1+\theta^*)} \\ \sqrt{1/(1+\theta^*)} & 1 \end{vmatrix}.$$

Since $\theta^* > 1$, we have immediately that the principal minors of T are positive. Hence T is positive definite, and the proof is complete.

Setting the r.h.s. of (5.11) equal to P^*, we obtain an approximation to the asymptotic sample size

$$(5.15) \qquad \sqrt{n} = -2\Phi^{-1}((1 - P^*)/(k - 1))/\log \theta^*.$$

6. Directions for future research

Several interesting problems are suggested by the results of this thesis. We pose some of these problems in this section.

Throughout Part II we used the Bonferroni and Slepian inequalities to find lower bounds on the PCS_a and then from these lower bounds we obtained approximations to the sample size. It would be interesting to compare these lower bounds on the PCS_a with approximations of the PCS_a (obtained by numerical integration) to get some idea of the "efficiency" of the inequalities in the problems we have considered. Of particular interest would be a comparison of the infimum (obtained by numerical search techniques) of this approximation of the PCS_a over the region of preference for a correct selection with the analytical results obtained from the inequalities.

A simpler problem which is also of interest is a comparison of the lower limit given in Theorem 4.3 with the infimum (obtained by numerical search techniques) of the lower bound of Theorem 4.2.

All of the results in Part II have been obtained using asymptotic distribution theory. It would be interesting to compare the PCS_a obtained from this theory with Monte Carlo estimates of the PCS. Of particular interest is the

comparison of these estimates of the PCS with the PCS_a as a function N and θ^*, when the parameters are in the LFC.

The sensitivity of the results of this thesis to the assumption that the distribution of $(x_0, x_1, \ldots, x_k)$ is multivariate normal is another problem that could be studied using Monte Carlo techniques.

The results of Sections 4.2 and 5.3 could be strengthened if the Slepian inequality could be used in place of the Bonferroni inequality. Although the conditions required for the Slepian inequality (Appendix B) appear to be satisfied for these more general cases, we have been unable to prove this analytically. The method employed in Section 4.4.4 to obtain this result appears to be too complicated and messy for these more general cases and hence a new approach is needed.

An obvious but seemingly difficult problem is the extension of the results of Chapter 5 for t = 1 to arbitrary t or at least to t = 2. The major difficulty here is the complicated nature of covariances of the sample conditional variances.

In all of the problem formulations we have considered in this thesis, the value of t was fixed prior to experimentation. In some situations it might be more reasonable to allow the value of t to be determined on the basis of the values of the observations. There are a number of ways in

which this new problem could be formulated. However, it appears that the same difficulties encountered in the formulations considered in this thesis would also be encountered in these new formulations.

Appendix A

Asymptotic joint distributions of the sample conditional variances

In this appendix, we find the asymptotic joint distributions of the sample statistics which are used in the decision procedures of Part II. All of these asymptotic distributions are multivariate normal and hence are defined by their expected values, variances, and covariances. We will denote these moments of the asymptotic distribution by E_a, Var_a, and Cov_a, respectively. (For each of the cases we consider, these quantities are equivalent to the corresponding asymptotic moments of the exact joint distribution, although this is, of course, not true in general.)

These asymptotic joint distributions are derived from the following two theorems. Theorem A.1 is an immediate result of Anderson's [2] Theorem 4.2.5 and Theorem A.2 is a generalization of his Theorem 4.2.6 as given by Rao [44], Section 6a.2 result iii.

We let n = N-t-1 and define

$$c_{ij} = v_{ij}/n\sqrt{\sigma_{ii}\sigma_{jj}}, \tag{A.1}$$

where v_{ij} is the $ij^{\underline{th}}$ element of the cross product matrix given by (3.10).

Theorem A.1

The asymptotic $((k+2)(k+1)/2)$-variate normal distribution of $\sqrt{n}\, c_{ij}$ $(i \leq j;\ i,j = 0, 1, \ldots, k)$ is determined by

$$E_a(\sqrt{n}\, c_{ij}) = \begin{cases} 1, & i=j \\ \rho_{ij}, & i \neq j \end{cases},$$

$$\mathrm{Var}_a(\sqrt{n}\, c_{ij}) = 1 + \rho_{ij}^2,$$

and

$$\mathrm{Cov}_a(\sqrt{n}\, c_{ij}, \sqrt{n}\, c_{rs}) = \rho_{ir}\rho_{js} + \rho_{is}\rho_{jr}.$$

Theorem A.2

If $\sqrt{n}\, \underline{u}$ has a p-variate asymptotic normal distribution determined by

$$E_a(\sqrt{n}\, u_i) = b_i,$$

$$\mathrm{Var}_a(\sqrt{n}\, u_i) = a_{ii},$$

$$\mathrm{Cov}_a(\sqrt{n}\, u_i, \sqrt{n}\, u_j) = a_{ij},$$

and $\underline{w} = \underline{f}(\underline{u})$ is a vector-valued function of the p-vector $\underline{u}$ such that each element of $\underline{w}$ is totally differentiable, then $\sqrt{n}\, \underline{w}$ has an asymptotic multivariate normal distribution determined by

$$E_a(\sqrt{n}\, w_i) = f_i(\underline{b}),$$

$$Var_a(\sqrt{n}\, w_i) = \underline{h}_i\, A\, \underline{h}_i',$$

and

$$Cov_a(\sqrt{n}\, w_i,\ \sqrt{n}\, w_j) = \underline{h}_i\, A\, \underline{h}_j'$$

where f_i is the $i^{\underline{th}}$ component of $\underline{f}$, A is the p by p covariance matrix of $\underline{u}$, and the $\underline{h}_i$ are p-vectors with elements given by

$$h_{ij} = \partial f_i/\partial u_j\big|_{\underline{u}=\underline{b}} \qquad (j=1,2,\ldots,p).$$

A.1 $\underline{t = 1}$

In this section we find the asymptotic joint distribution of the

$$y_i' = \sqrt{n}(\log(s^2_{0.(1)}) - \log(s^2_{0.(k-i+1)})) \qquad (i=1,2,\ldots,k-1). \tag{A.2}$$

To obtain this result, we first find the joint asymptotic distribution of the $s^2_{0.i}$. Using the notation of the last section, we have

$$s^2_{0.i} = \sigma_{00}(c_{00} - c^2_{0i}/c_{ii}). \tag{A.3}$$

Applying Theorem A.2 to Theorem A.1 and (A.3) we obtain the asymptotic distribution of the $\sqrt{n}\, s^2_{0.i}$.

Corollary A.1

The asymptotic k-variate normal distribution of the $\sqrt{n}\ s^2_{0.i}$ is determined by

$$E_a(\sqrt{n}\ s^2_{0.i}) = \sigma_{00}(1 - \rho^2_{0i}),$$

$$Var_a(\sqrt{n}\ s^2_{0.i}) = 2\sigma^2_{00}(1 - \rho^2_{0i})^2,$$

and

$$Cov_a(\sqrt{n}\ s^2_{0.i}, \sqrt{n}\ s^2_{0.j}) =$$

$$2\sigma^2_{00}(1 - \rho^2_{0i} - \rho^2_{0j} + \rho_{0i}\rho_{0j}\rho_{ij})^2 \qquad (i \neq j).$$

Applying Theorem A.2 to this result, we obtain the asymptotic joint distribution of the $\sqrt{n}\ \log\ s^2_{0.i}$.

Corollary A.2

The asymptotic k-variate normal distribution of the $\sqrt{n}\ \log\ s^2_{0.i}$ is determined by

$$E_a(\sqrt{n}\ \log\ s^2_{0.i}) = \log\ \sigma_{00}(1 - \rho^2_{0i}),$$

$$Var_a(\sqrt{n}\ \log\ s^2_{0.i}) = 2,$$

and

$$\mathrm{Cov}_a(\sqrt{n}\,\log s^2_{0.i},\ \sqrt{n}\,\log s^2_{0.j}) =$$

$$\frac{2(1 - \rho^2_{0i} - \rho^2_{0j} + \rho_{0i}\rho_{0j}\rho_{ij})^2}{(1 - \rho^2_{0i})(1 - \rho^2_{0j})} \qquad (i \neq j).$$

From this result using the notation of Section 3.4 we obtain the corresponding result for the $\sqrt{n}\,\log s^2_{0.(i)}$.

Corollary A.3

The asymptotic k-variate normal distribution of the $\sqrt{n}\,\log s^2_{0.(i)}$ is determined by

$$E_a(\sqrt{n}\,\log s^2_{0.(k-i+1)}) = \log(\sigma_{00}(1 - \rho^2_{0[i]})),$$

$$\mathrm{Var}_a(\sqrt{n}\,\log s^2_{0.(k-i+1)}) = 2,$$

and

$$\mathrm{Cov}_a(\sqrt{n}\,\log s^2_{0.(k-i+1)},\ \sqrt{n}\,\log s^2_{0.(k-j+1)}) = 2\omega_{ij} \qquad (i \neq j),$$

where

$$\omega_{ij} = \frac{(1 - \rho^2_{0[i]} - \rho^2_{0[j]} + \rho_{0[i]}\rho_{0[j]}\rho_{[i][j]})^2}{(1 - \rho^2_{0[i]})(1 - \rho^2_{0[j]})} \qquad (i \neq j),$$

and the $\rho_{[i][j]}$ are defined in Section 5.1.

Using this result we again apply Theorem A.2 to obtain

the following corollary which is used in Section 5.1 to express the PCS_a in terms of a standard multivariate normal distribution function.

Corollary A.4

The asymptotic (k-1)-variate normal distribution of the

$$y_i = \frac{\sqrt{n}\,\log(s^2_{0.(1)}/s^2_{0.(k-i+1)})}{2\sqrt{1-\omega_{ik}}} \qquad (i=1,2,\ldots,k-1)$$

is determined by

$$E_a(y_i) = \frac{\log((1-\rho^2_{0[i]})/(1-\rho^2_{0[k]}))}{2\sqrt{1-\omega_{jk}}},$$

$$Var_a(y_i) = 1,$$

and

$$Cov_a(y_i, y_j) = \omega'_{ij} \qquad (i\neq j),$$

where

$$\omega'_{ij} = \frac{(1-\omega_{ik}-\omega_{jk}+\omega_{ij})}{2\sqrt{(1-\omega_{ik})(1-\omega_{jk})}} \qquad (i\neq j).$$

For the case of uncorrelated predictor variates $\rho_{[i][j]} = 0$ $(i \neq j;\ i,j = 1, 2, \ldots, k)$, and we have the following corollary which is used in Section 4.1.2 to give an

expression for the PCS_a (4.20).

Corollary A.4a

The asymptotic (k-1)-variate distribution of the

$$(A.4) \qquad y_i = \frac{\sqrt{n}\,\log(s^2_{0.(1)}/s^2_{0.(k-i+1)})}{2\sqrt{1-\gamma_{ik}}} \qquad (i=1,2,\ldots,k-1)$$

is determined by

$$E_a(y_i) = \frac{\log((1-\rho^2_{0[i]})/(1-\rho^2_{0[k]}))}{2\sqrt{1-\gamma_{ik}}},$$

$$Var_a(y_i) = 1,$$

and

$$Cov_a(y_i, y_j) = \gamma'_{ij} \qquad (i\neq j),$$

where

$$\gamma'_{ij} = \frac{(1-\gamma_{ik}-\gamma_{jk}+\gamma_{ij})}{2\sqrt{(1-\gamma_{ik})(1-\gamma_{jk})}} \qquad (i\neq j)$$

and

$$\gamma_{ij} = \frac{(1-\rho^2_{0[i]}-\rho^2_{0[j]})^2}{(1-\rho^2_{0[i]})(1-\rho^2_{0[j]})} \qquad (i\neq j).$$

A.2 t > 1, uncorrelated predictor variates

In this section we find the asymptotic joint distribution of the sample conditional variances, each of which is based on t uncorrelated predictor variates. Again our objective is to find the joint distribution of the y_i' (A.2), and we proceed in a manner similar to that used in the previous section by first finding the asymptotic joint distribution of the $s^2_{0.\alpha}$.

The following corollary is obtained using Theorems A.1 and A.2, after writing the $s^2_{0.\alpha}$ in terms of the c_{ij} in a manner similar to that of (A.3). The expressions for the $s^2_{0.\alpha}$ are, of course, considerably more complicated than (A.3).

Corollary A.5

The asymptotic U-variate normal distribution of the $\sqrt{n}\ s^2_{0.\alpha}$ is determined by

$$E_a(\sqrt{n}\ s^2_{0.\alpha}) = \sigma_{00}(1 - \bar{R}^2_{0.\alpha})$$

$$Var_a(\sqrt{n}\ s^2_{0.\alpha}) = 2\sigma^2_{00}(1 - \bar{R}^2_{0.\alpha})^2,$$

and

$$Cov_a(\sqrt{n}\ s^2_{0.\alpha}, \sqrt{n}\ s^2_{0.\alpha''}) = 2\sigma^2_{00}(1 - \bar{R}^2_{0.\alpha^*})^2 \qquad (\alpha \neq \alpha''),$$

where α^*, u_α, and $\bar{R}_{0.\alpha^*}$ are defined in Section 4.2.1.

We apply Theorem A.2 to this result to obtain the following:

Corollary A.6

The asymptotic U-variate normal distribution of the $\sqrt{n}\ \log s^2_{0.\alpha}$ is determined by

$$E_a(\sqrt{n}\ \log s^2_{0.\alpha}) = \log(\sigma_{00}(1 - \bar{R}^2_{0.\alpha})),$$

$$Var_a(\sqrt{n}\ \log s^2_{0.\alpha}) = 2,$$

and

$$Cov_a(\sqrt{n}\ \log s^2_{0.\alpha}, \sqrt{n}\ \log s^2_{0.\alpha''}) =$$

$$\frac{2(1 - \bar{R}^2_{0.\alpha^*})^2}{(1 - \bar{R}^2_{0.\alpha})(1 - \bar{R}^2_{0.\alpha''})} \qquad (\alpha \neq \alpha'').$$

From this result using the notation of Section 4.2.1, we obtain the corresponding result for the $\sqrt{n}\ \log s^2_{0.(i)}$.

Corollary A.7

The asymptotic U-variate distribution of the $\sqrt{n}\ \log s^2_{0.(i)}$ is determined by

$$E_a(\sqrt{n}\ \log s^2_{0.(i)}) = \log(\sigma_{00}(1 - \bar{R}^2_{0.[i]})),$$

$$Var_a(\sqrt{n}\ \log s^2_{0.(i)}) = 2,$$

and

$$Cov_a(\sqrt{n}\ \log s^2_{0.(i)}, \sqrt{n}\ \log s^2_{0.(j)}) = 2\lambda_{ij} \qquad (i \neq j),$$

where

$$\lambda_{ij} = \frac{(1 - \sum_{m}^{i,j} \rho_{0m}^2)^2}{(1 - \bar{R}^2_{0.[i]})(1 - \bar{R}^2_{0.[j]})} \qquad (i \neq j)$$

and the summation notation is defined in Section 4.2.1.

Using this result, we again apply Theorem A.2 to obtain the following corollary, which is used in Section 4.2.2 to express the PCS_a (4.58) in terms of a standard multivariate normal distribution function.

Corollary A.8

The asymptotic (U-1)-variate normal distribution of the

$$(A.5) \qquad y_i = \frac{\sqrt{n}\, \log(s^2_{0.(1)}/s^2_{0.(k-i+1)})}{2\sqrt{1 - \lambda_{ik}}} \qquad (i=1,2,\ldots,U-1)$$

is determined by

$$E_a(y_i) = \frac{\log((1 - \bar{R}^2_{0.[i]})/(1 - \bar{R}^2_{0.[k]}))}{2\sqrt{1 - \lambda_{ik}}}$$

$$Var_a(y_i) = 1,$$

and

$$Cov_a(y_i, y_j) = \lambda'_{ij} \qquad (i \neq j),$$

where

$$\lambda'_{ij} = \frac{(1 - \lambda_{ik} - \lambda_{jk} + \lambda_{ij})}{2\sqrt{(1 - \lambda_{ik})(1 - \lambda_{jk})}} \qquad (i \neq j).$$

Appendix B

The Bonferroni and the Slepian inequalities

After stating the Bonferroni and the Slepian inequalities as lemmas, we use each of them to obtain lower bounds on multivariate normal probabilities. These results are used in Part II to give lower bounds on the expressions for the PCS_a.

First we give the Bonferroni inequality (e.g., Feller [18], p. 100).

Lemma B.1

Let $A_1, A_2, \ldots, A_p$ denote a sequence of events and let A_i' denote the complement of the event A_i. Then

$$P\{\bigcup_{i=1}^{p} A_i\} \leq \sum_{i=1}^{p} P\{A_i\}$$

and hence

$$P\{\bigcap_{i=1}^{p} A_i\} = 1 - P\{\bigcup_{i=1}^{p} A_i'\}$$

$$\geq 1 - \sum_{i=1}^{p} \{A_i'\}$$

If we define the events

$$A_i = [y_i \leq a_i] \qquad (i=1,2,\ldots,p),$$

then by this lemma, we have

$$\text{(B.1)} \qquad \Phi_p(a_1, a_2, \ldots, a_p) \geq 1 - \sum_{i=1}^{p} (1 - \Phi(a_i)),$$

where, as before, Φ_p denotes a standard multivariate normal distribution function having zero means and unit variances.

The following lemma is due to Slepian [48].

Lemma B.2

Let Φ_p^{ω} and Φ_p^{κ} denote p-variate normal distribution functions with zero expectations and nonnegative definite covariance matrices given by $\{\omega_{ij}\}$ and $\{\kappa_{ij}\}$ respectively. If

$$\omega_{ij} \geq \kappa_{ij} \qquad (i \neq j; i,j = 1,2,\ldots,p)$$

and

$$\omega_{ii} = \kappa_{ii} \qquad (i = 1,2,\ldots,p),$$

then

$$\Phi_p^{\kappa}(a_1, a_2, \ldots, a_p) \geq \Phi_p^{\kappa}(a_1, a_2, \ldots, a_p).$$

Using this lemma, we have

$$\text{(B.2)} \qquad \Phi_p^{\omega}(a_1, a_2, \ldots, a_p) \geq \prod_{i=1}^{p} \Phi(a_i),$$

whenever

$$\omega_{ii} = 1 \qquad (i=1,2,\ldots,p)$$

and

$$\omega_{ij} \geq 0 \qquad (i\neq j; i,j=1,2,\ldots,p)$$

In the following lemma, we show that the Slepian inequality (when its assumptions are satisfied) gives a better bound than does the Bonferroni inequality.

Lemma B.3

$$\text{(B.3)} \qquad \prod_{i=1}^{p} \Phi(a_i) \geq 1 - \sum_{i=1}^{p} (1-\Phi(a_i)).$$

Proof:

This result follows directly from the Bonferroni inequality by letting $a_1, a_2, \ldots, a_p$ be a sequence of events and denoting the probability of the event a_i by $\Phi(a_i)$.

Appendix C

A numerical comparison of the Bonferroni and the Slepian inequalities

In this appendix we study the efficiency of the Bonferroni and the Slepian inequalities by using each of them to approximate the sample size for a problem in which the exact sample size is known. The problem is that of selecting the univariate normal population with the largest population mean, when each of the populations has a common known variance (σ^2). The PCS for this problem can be expressed (Bechhofer [5]) in terms of a (k-1)-variate standard normal distribution function with zero means, unit variances, and off-diagonal covariances of 1/2.

Using the Bonferroni and the Slepian inequalities to give lower bounds on this PCS expression, we obtain two conservative approximations to the sample size denoted by N_B and N_S, respectively, where

$$\text{(C.1)} \qquad \sqrt{N}_B = 2(\sigma/\delta^*)\Phi^{-1}((1 - P^*)/(k - 1)),$$

$$\text{(C.2)} \qquad \sqrt{N}_S = 2(\sigma/\delta^*)\Phi^{-1}(P^{*1/(k-1)})$$

and $\{\delta^*, P^*\}$ are the preassigned constants.

The exact sample size N can be computed by

(C.3) $\sqrt{N} = (\sigma/\delta^*)d(P^*, k)$,

where the values of d are given in Table 1 of Bechhofer [5].

The values of N_B/N and N_S/N are given in Table C.1 for k = 3, 6, 9 and P* values ranging from .5 to .999.

For high P* ($P^* \geq .99$), both of the inequalities provide good approximations to the exact sample size (within 5% for $k \leq 9$), and the N_B is only slightly larger than N_S.

For $P^* < .9$, neither of the inequalities provides a good approximation to the sample size (N_B and N_S are both at least 15% greater than N for the values of k considered), and N_B/N is considerably larger than N_S/N.

For $.9 \leq P^* < .99$, the Slepian inequality provides a reasonable approximation when k is not too large, but the approximation provided by the Bonferroni inequality is not as good.

Table C.1

A Comparison of Sample Approximations Derived from the Bonferroni and the Slepian Inequalities with the Exact Requirements for the Problem of Ranking Normal Means

k =	3		6		9	
P*	N_B/N	N_S/N	N_B/N	N_S/N	N_B/N	N_S/N
.5	2.93529	1.91766	2.47320	1.91914	2.35754	1.92193
.6	1.80715	1.45038	1.85936	1.58701	1.85087	1.62104
.7	1.40178	1.25527	1.51766	1.38337	1.54029	1.42128
.8	1.20335	1.14557	1.30124	1.24123	1.32900	1.27371
.9	1.08838	1.07172	1.14913	1.12988	1.17077	1.15237
.95	1.04652	1.04070	1.08496	1.07782	1.10037	1.09340
.99	1.01444	1.01376	1.02948	1.02857	1.03660	1.03568
.995	1.00941	1.00912	1.01980	1.01941	1.02500	1.02460
.999	1.00381	1.00376	1.00846	1.00839	1.01101	1.01089

Bibliography

1. Alam, K. and Rizvi, M. H : "Selection from multivariate normal populations," Annals of the Institute of Statistical Mathematics (Tokyo), Vol 18, No. 3 (1966), pp. 307-318.

2. Anderson, T. W.: Introduction to Multivariate Statistical Analysis, John Wiley & Sons, Inc., 1958.

3. Barr, D. R. and Rizvi, M. H.: "An introduction to ranking and selection procedures," Journal of the American Statistical Association, Vol. 61 (1966), pp. 640-646.

4. Bartlett, M. S. and Kendall, D. G.: "The statistical analysis of variance-heterogeneity and the logarithmic transformation," Journal of the Royal Statistical Society Supplement, Vol. 8 (1946), pp. 128-138.

5. Bechhofer, R. E.: "A single-sample multiple-decision procedure for ranking means of normal populations with known variances," Annals of Mathematical Statistics, Vol. 25 (1954), pp. 16-39.

6. Bechhofer, R. E.: "A multiplicative model for analyzing variances which are affected by several factors," Journal of the American Statistical Association, Vol. 55 (1960), pp. 245-264.

7. Bechhofer, R. E., Dunnett, Charles, and Sobel, Milton: "A two-sample multiple decision procedure for ranking means of normal populations with a common unknown variance," Biometrika, Vol. 41 (1954), Parts 1 and 2, pp. 170-176.

8. Bechhofer, R. E., Elmaghraby, S. A., and Morse, N.: "A single-sample multiple-decision procedure for selecting the multinomial event which has the largest probability," Annals of Mathematical Statistics, Vol. 30 (1959), pp. 102-119.

9. Bechhofer, R. E., Kiefer, J., and Sobel, M.: Sequential Identification and Ranking Procedures, Statistical Research Monographs, Vol. 3, The University of Chicago Press, 1968.

10. Bechhofer, R. E. and Sobel, M.: "A single-sample multiple-decision procedure for ranking variances of normal populations," Annals of Mathematical Statistics, Vol. 25 (1954), pp. 273-289.

11. Box, G. E. P.: "Non-normality and tests on variances," Biometrika, Vol. 40 (1953), Parts 1 and 2, pp. 318-335.

12. Cramer, H.: Mathematical Methods of Statistics, Princeton University Press, 1946.

13. Chew, V.: "Confidence, prediction, and tolerance regions for the multivariate normal distribution," Journal of the American Statistical Association, Vol. 61 (1961), pp. 605-617.

14. Draper, N. R. and Smith, H.: Applied Regression Analysis, John Wiley & Sons, Inc., 1966.

15. Eaton, M. L.: "Some optimum properties of ranking procedures," Annals of Mathematical Statistics, Vol. 38 (1967), pp. 124-137.

16. Eaton, M. L.: "The generalized variance: testing and ranking problem," Annals of Mathematical Statistics, Vol. 38 (1967), pp. 941-943.

17. Fairweather, W. R.: "Some extensions of Somerville's procedure for ranking means of normal populations," Abstract, Annals of Mathematical Statistics, Vol. 38 (1967), p. 961.

18. Feller, W.: An Introduction to Probability Theory and Its Applications, Vol. 1, 2nd Ed., John Wiley & Sons, Inc., 1957.

19. Ferguson, T. S.: Mathematical Statistics: A Decision Theoretic Approach, Academic Press, Inc., 1967.

20. Fisz, M.: Probability Theory and Mathematical Statistics, 3rd Ed., John Wiley & Sons, Inc., 1963.

21. Garside, M. J.: "The best subset in multiple regression analysis," Applied Statistics Journal of the Royal Statistical Society, Series C, Vol. 15 (1965), pp. 196-200.

22. Geary, R. C.: "Ex post determination of significance in multivariate regression when the independent variables are orthogonal," Journal of the Royal Statistical Society, Series B, Vol. 29, No. 1 (1967), pp. 154-161.

23. Gnanadesikan, M. R.: "Some selection and ranking procedures for multivariate normal populations," Abstract, Annals of Mathematical Statistics, Vol. 37 (1966) p. 1418.

24. Graybill, F.: An Introduction to Linear Statistical Models, Vol. 1, McGraw-Hill Book Company, 1961.

25. Gupta, S. S.: "Probability integrals of the multivariate normal and multivariate t," Annals of Mathematical Statistics, Vol. 34 (1963), pp. 792-828.

26. Gupta, S. S.: "Bibliography of the multivariate normal integrals and related topics," Annals of Mathematical Statistics, Vol. 34 (1963), pp. 829-838.

27. Gupta, S. S.: "On some multiple decision (selection and ranking) rules," Technometrics, Vol. 7, No. 2 (1965), pp. 225-246.

28. Gupta, S. S.: "On some selection and ranking procedures with application to multivariate populations," Multivariate Analysis: Proceedings of an International Symposium, edited by P. R. Krishnaiah, Academic Press, Inc., 1966, pp. 457-475.

29. Gupta, S. S. and Panchapekesan, S.: "Some selection and ranking procedures for multivariate normal populations," Mimeograph Series No. 163, Department of Statistics, Purdue University, Lafayette, Indiana, 1968.

30. Gupta, S. S. and Studden, W. J.: "On some selection and ranking procedures with applications to multivariate populations," Mimeograph Series No. 58, Department of Statistics, Purdue University, Lafayette, Indiana, 1965.

31. Hall, W. J.: "Most economical multiple decision rules," Annals of Mathematical Statistics, Vol. 29 (1958), pp. 1079-1094.

32. Hall, W. J.: "The most economical character of some Bechhofer and Sobel decision rules," Annals of Mathematical Statistics, Vol. 30 (1959), pp. 964-969.

33. Hocking, R. R. and Leslie, R. N.: "Selection of the best subset in regression analysis," Mimeo, Texas A & M University (1967).

34. Hotelling, H.: "The selection of variates for use in prediction with some comments on the general problem of nuisance parameters," Annals of Mathematical Statistics, Vol. 11 (1940), pp. 271-283.

35. Kesten, H. and Morse, N.: "A property of the multinomial distribution," Annals of Mathematical Statistics, Vol. 30 (1959), pp. 120-127.

36. Krishnaiah, P. R.: "Selection procedures based on covariance matrices of multivariate normal populations," Blanch Anniversary Volume, Aerospace Research Laboratories, Office of Aerospace Research, United States Air Force, 1967, pp. 149-160.

37. Krishnaiah, P. R. and Rizvi, M. H.: "Some procedures for selection of the multivariate normal populations better than a control," Multivariate Analysis: Proceedings of an International Symposium, edited by P. R. Krishnaiah, Academic Press, 1966.

38. Mallows, C. L.: "Choosing a subset regression," Mimeo, Bell Telephone Laboratories, 1967.

39. Miller, K. S.: Multidimensional Gaussian Distributions, John Wiley and Sons, Inc., 1964.

40. Milton, R. C.: "Tables of the equally correlated multivariate normal probability integral," Technical Report No. 27, Department of Statistics, University of Minnesota, Minneapolis, Minnesota, 1963.

41. Olkin, I.: Personal communication.

42. Paulson, E.: "A sequential procedure for selecting the population with the largest mean from k normal populations," Annals of Mathematical Statistics, Vol. 35 (1964), pp. 174-180.

43. Plackett, R. L.: "A reduction formula for normal multivariate integrals," Biometrika, Vol. 41 (1954), pp. 351-360.

44. Rao, C. R.: Linear Statistical Inference and Its Application, John Wiley & Sons, Inc., 1965.

45. Rizvi, M. H.: "Ranking and selection problems of normal populations using the absolute values of their means," fixed sample size case, Technical Report No. 31, Department of Statistics, University of Minnesota, Minneapolis, Minnesota, 1963.

46. Robbins, H., Sobel, M., and Starr, N.: "A sequential procedure for selecting the largest of k means," Annals of Mathematical Statistics, Vol. 39 (1968), p. 88-92.

47. Sidak, Z.: "Rectangular confidence regions for the means of multivariate normal distributions," Journal of the American Statistical Association, Vol. 62 (1967), pp. 626-633.

48. Slepian, D.: "The one-sided barrier problem for Gaussian Noise," The Bell System Technical Journal, Vol. 41, No. 2 (1962), pp. 463-502.

49. Thornby, J. I.: "A study of the ranking criteria for selecting a subset containing the best of k bivariate normal populations," Technical Report No. 36, Department of Statistics, University of Minnesota, Minneapolis, Minnesota, 1964.

Unclassified
Security Classification

DOCUMENT CONTROL DATA - R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1 ORIGINATING ACTIVITY *(Corporate author)*
Department of Operations Research
College of Engineering, Cornell University
Ithaca, New York 14850

2a REPORT SECURITY CLASSIFICATION
Unclassified

2b GROUP

3 REPORT TITLE

A MULTIPLE-DECISION APPROACH TO THE SELECTION OF THE BEST SET OF PREDICTOR VARIATES

4 DESCRIPTIVE NOTES *(Type of report and inclusive dates)*
Technical Report, July 1969

5 AUTHOR(S) *(Last name, first name, initial)*

John Schmidt Ramberg

6 REPORT DATE	7a TOTAL NO OF PAGES	7b NO OF REFS
July 1969	102 + 3	49

8a CONTRACT OR GRANT NO.
DA-31-124-ARO-D-474

b [redacted]

c Nonr-401(53)

d

9a ORIGINATOR'S REPORT NUMBER(S)
Technical Report No. 79

9b OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*

10 AVAILABILITY/LIMITATION NOTICES

Distribution of this document is unlimited

11 SUPPLEMENTARY NOTES
Sponsoring military activity
U.S. Army Research Office
Durham, N.C. 27706

12 SPONSORING MILITARY ACTIVITY
Logistics and Mathematical Statistics
Branch, Office of Naval Research
Washington, D.C. 20360

13 ABSTRACT

Some "indifference zone" multiple-decision selection procedure formulations of prediction problems involving multivariate normal populations are considered. These problems are of two types. In Part I we consider problems involving k bivariate normal populations, where the goal is to select the "best" population. In this part the "goodness" of the prediction is measured in terms of three different parameters -- the population conditional variance, the population correlation coefficient, and the absolute value of the population correlation coefficent.

In Part II we consider the problem of selecting the best set of a preassigned number t variates from a set of k predictor variates for predicting a designated variate, the predictand. The "best" set of predictor variates is defined to be the set of t variates for which the predictand has the smallest population conditional variance (or equivalently the largest population multiple correlation coefficient). Sample size requirements are obtained using asymptotic distribution theory of the transformed statistics. (Monte Carlo sampling studies indicate these results are valid for relatively small sample sizes.) For $k > 2$, the Bonferroni and Slepian inequalities are used to obtain conservative sample size approximations.

DD FORM 1 JAN 64 1473

Unclassified
Security Classification

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Correlation coefficient Decision theory Mathematical statistics Multivariate prediction Prediction theory Ranking procedures Selection procedures						

INSTRUCTIONS

1. ORIGINATING ACTIVITY: Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization *(corporate author)* issuing the report.

2a. REPORT SECURITY CLASSIFICATION: Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2b. GROUP: Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. REPORT TITLE: Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. DESCRIPTIVE NOTES: If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. AUTHOR(S): Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. REPORT DATE: Enter the date of the report as day, month, year; or month, year. If more than one date appears on the report, use date of publication.

7a. TOTAL NUMBER OF PAGES: The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7b. NUMBER OF REFERENCES: Enter the total number of references cited in the report.

8a. CONTRACT OR GRANT NUMBER: If appropriate, enter the applicable number of the contract or grant under which the report was written.

8b, 8c, & 8d. PROJECT NUMBER: Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9a. ORIGINATOR'S REPORT NUMBER(S): Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9b. OTHER REPORT NUMBER(S): If the report has been assigned any other report numbers *(either by the originator or by the sponsor)*, also enter this number(s).

10. AVAILABILITY/LIMITATION NOTICES: Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through _____."

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through _____."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through _____."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. SUPPLEMENTARY NOTES: Use for additional explanatory notes.

12. SPONSORING MILITARY ACTIVITY: Enter the name of the departmental project office or laboratory sponsoring *(paying for)* the research and development. Include address.

13. ABSTRACT: Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as *(TS), (S), (C), or (U)*.

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. KEY WORDS: Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, rules, and weights is optional.